Quantum Wells, Wires and Dots

Quantum Wells, Wires and Dots

Theoretical and Computational Physics of Semiconductor Nanostructures

Fourth Edition

Paul Harrison

Sheffield Hallam University, UK

Alex Valavanis

The University of Leeds, UK

WILEY

This edition first published 2016
© 2016 John Wiley & Sons, Ltd

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the authors to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and authors have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the authors shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

The advice and strategies contained herein may not be suitable for every situation. In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of experimental reagents, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each chemical, piece of equipment, reagent, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. The fact that an organization or website is referred to in this work as a citation and/or a potential source of further information does not mean that the authors or the publisher endorses the information the organization or website may provide or recommendations it may make. Further, readers should be aware that Internet websites listed in this work may have changed or disappeared between when this work was written and when it is read. No warranty may be created or extended by any promotional statements for this work. Neither the publisher nor the authors shall be liable for any damages arising herefrom.

Library of Congress Cataloging-in-Publication Data

Names: Harrison, P. (Paul), author. | Valavanis, Alex, author.
Title: Quantum wells, wires and dots : theoretical and computational physics of semiconductor nanostructures / Paul Harrison (Sheffield Hallam University, UK), Alex Valavanis (The University of Leeds, UK).
Description: Fourth edition. | West Sussex, United Kingdom ; Hoboken, NJ : John Wiley & Sons, Inc., 2016. | 2016 | Includes bibliographical references and index.
Identifiers: LCCN 2015038317 (print) | LCCN 2015045402 (ebook) | ISBN 9781118923368 (cloth) | ISBN 1118923367 (cloth) | ISBN 9781118923351 (pdf) | ISBN 9781118923344 (epub)
Subjects: LCSH: Quantum wells. | Nanowires. | Quantum dots.
Classification: LCC QC176.8.Q35 H37 2016 (print) | LCC QC176.8.Q35 (ebook) | DDC 537.6/226–dc23
LC record available at http://lccn.loc.gov/2015038317

A catalogue record for this book is available from the British Library.

ISBN: 9781118923368

1 2016

Dedication

To our families

Contents

List of contributors **xv**

Preface **xvii**

Acknowledgements **xxi**

Introduction **xxv**
References xxvi

1 Semiconductors and heterostructures **1**
1.1 The mechanics of waves 1
1.2 Crystal structure 3
1.3 The effective mass approximation 5
1.4 Band theory 5
1.5 Heterojunctions 7
1.6 Heterostructures 8
1.7 The envelope function approximation 10
1.8 Band non-parabolicity 11
1.9 The reciprocal lattice 13
Exercises 16
References 17

2 Solutions to Schrödinger's equation **19**
2.1 The infinite well 19
2.2 In-plane dispersion 22
2.3 Extension to include band non-parabolicity 24
2.4 Density of states 26
2.4.1 Density-of-states effective mass 28
2.4.2 Two-dimensional systems 29
2.5 Subband populations 32
2.5.1 Populations in non-parabolic subbands 34
2.5.2 Calculation of quasi-Fermi energy 35
2.6 Thermalised distributions 36

2.7 Finite well with constant mass . . . 38
2.7.1 Unbound states . . . 44
2.7.2 Effective mass mismatch at heterojunctions . . . 46
2.7.3 The infinite barrier height and mass limits . . . 49
2.8 Extension to multiple-well systems . . . 51
2.9 The asymmetric single quantum well . . . 53
2.10 Addition of an electric field . . . 55
2.11 The infinite superlattice . . . 58
2.12 The single barrier . . . 64
2.13 The double barrier . . . 66
2.14 Extension to include electric field . . . 72
2.15 Magnetic fields and Landau quantisation . . . 73
2.16 In summary . . . 75
Exercises . . . 75
References . . . 77

3 Numerical solutions **81**
3.1 Bisection root-finding . . . 81
3.2 Newton–Raphson root finding . . . 83
3.3 Numerical differentiation . . . 85
3.4 Discretised Schrödinger equation . . . 86
3.5 Shooting method . . . 86
3.6 Generalised initial conditions . . . 88
3.7 Practical implementation of the shooting method . . . 90
3.8 Heterojunction boundary conditions . . . 93
3.9 Matrix solutions of the discretised Schrödinger equation . . . 93
3.10 The parabolic potential well . . . 96
3.11 The Pöschl–Teller potential hole . . . 100
3.12 Convergence tests . . . 101
3.13 Extension to variable effective mass . . . 102
3.14 The double quantum well . . . 105
3.15 Multiple quantum wells and finite superlattices . . . 108
3.16 Addition of electric field . . . 108
3.17 Extension to include variable permittivity . . . 109
3.18 Quantum-confined Stark effect . . . 111
3.19 Field-induced anti-crossings . . . 111
3.20 Symmetry and selection rules . . . 112
3.21 The Heisenberg uncertainty principle . . . 113
3.22 Extension to include band non-parabolicity . . . 115
3.23 Poisson's equation . . . 117
3.24 Matrix solution of Poisson's equation . . . 121
3.25 Self-consistent Schrödinger–Poisson solution . . . 121
3.26 Modulation doping . . . 124
3.27 The high-electron-mobility transistor . . . 124
3.28 Band filling . . . 125

Exercises . . . 126
References . . . 128

4 Diffusion **131**
4.1 Introduction . . . 131
4.2 Theory . . . 133
4.3 Boundary conditions . . . 135
4.4 Convergence tests . . . 135
4.5 Numerical stability . . . 137
4.6 Constant diffusion coefficients . . . 137
4.7 Concentration-dependent diffusion coefficient . . . 139
4.8 Depth-dependent diffusion coefficient . . . 140
4.9 Time-dependent diffusion coefficient . . . 142
4.10 δ-doped quantum wells . . . 143
4.11 Extension to higher dimensions . . . 145
Exercises . . . 146
References . . . 147

5 Impurities **149**
5.1 Donors and acceptors in bulk material . . . 149
5.2 Binding energy in a heterostructure . . . 151
5.3 Two-dimensional trial wave function . . . 156
5.4 Three-dimensional trial wave function . . . 163
5.5 Variable-symmetry trial wave function . . . 169
5.6 Inclusion of a central cell correction . . . 175
5.7 Special considerations for acceptors . . . 176
5.8 Effective mass and dielectric mismatch . . . 177
5.9 Band non-parabolicity . . . 178
5.10 Excited states . . . 178
5.11 Application to spin-flip Raman spectroscopy in diluted magnetic semiconductors . . . 179
5.11.1 Diluted magnetic semiconductors . . . 179
5.11.2 Spin-flip Raman spectroscopy . . . 180
5.12 Alternative approach to excited impurity states . . . 183
5.13 Direct evaluation of the expectation value of the Hamiltonian for the ground state . . . 185
5.14 Validation of the model for the position dependence of the impurity . . . 186
5.15 Excited states . . . 187
5.16 Impurity occupancy statistics . . . 190
Exercises . . . 194
References . . . 195

6 Excitons **197**
6.1 Excitons in bulk . . . 197
6.2 Excitons in heterostructures . . . 199

6.3 Exciton binding energies . . . 199
6.4 1s exciton . . . 204
6.5 The two-dimensional and three-dimensional limits . . . 208
6.6 Excitons in single quantum wells . . . 212
6.7 Excitons in multiple quantum wells . . . 214
6.8 Stark ladders . . . 216
6.9 Self-consistent effects . . . 218
6.10 2s exciton . . . 219
Exercises . . . 220
References . . . 221

7 Strained quantum wells 223
7.1 Stress and strain in bulk crystals . . . 223
7.2 Strain in quantum wells . . . 227
7.3 Critical thickness of layers . . . 230
7.4 Strain balancing . . . 232
7.5 Effect on the band profile of quantum wells . . . 234
7.6 The piezoelectric effect . . . 237
7.7 Induced piezoelectric fields in quantum wells . . . 240
7.8 Effect of piezoelectric fields on quantum wells . . . 242
Exercises . . . 245
References . . . 247

8 Simple models of quantum wires and dots 249
8.1 Further confinement . . . 249
8.2 Schrödinger's equation in quantum wires . . . 251
8.3 Infinitely deep rectangular wires . . . 253
8.4 Simple approximation to a finite rectangular wire . . . 255
8.5 Circular cross-section wire . . . 260
8.6 Quantum boxes . . . 263
8.7 Spherical quantum dots . . . 264
8.8 Non-zero angular momentum states . . . 267
8.9 Approaches to pyramidal dots . . . 270
8.10 Matrix approaches . . . 271
8.11 Finite-difference expansions . . . 271
8.12 Density of states . . . 273
Exercises . . . 275
References . . . 276

9 Quantum dots 279
9.1 Zero-dimensional systems and their experimental realisation . . . 279
9.2 Cuboidal dots . . . 281
9.3 Dots of arbitrary shape . . . 282
9.3.1 Convergence tests . . . 287
9.3.2 Efficiency . . . 289

9.3.3 Optimisation . . . 291
9.4 Application to real problems . . . 292
9.4.1 InAs/GaAs self-assembled quantum dots . . . 292
9.4.2 Working assumptions . . . 293
9.4.3 Results . . . 294
9.4.4 Concluding remarks . . . 297
9.5 A more complex model is not always a better model . . . 299
Exercises . . . 300
References . . . 301

10 Carrier scattering 303
10.1 Introduction . . . 303
10.2 Fermi's golden rule . . . 304
10.3 Extension to sinusoidal perturbations . . . 306
10.4 Averaging over two-dimensional carrier distributions . . . 307
10.5 Phonons . . . 308
10.6 Longitudinal optic phonon scattering of two-dimensional carriers . . . 311
10.7 Application to conduction subbands . . . 323
10.8 Mean intersubband longitudinal optic phonon scattering rate . . . 325
10.9 Ratio of emission to absorption . . . 326
10.10 Screening of the longitudinal optical phonon interaction . . . 328
10.11 Acoustic deformation potential scattering . . . 329
10.12 Application to conduction subbands . . . 334
10.13 Optical deformation potential scattering . . . 336
10.14 Confined and interface phonon modes . . . 338
10.15 Carrier–carrier scattering . . . 339
10.16 Addition of screening . . . 348
10.17 Mean intersubband carrier–carrier scattering rate . . . 350
10.18 Computational implementation . . . 351
10.19 Intrasubband versus intersubband . . . 353
10.20 Thermalised distributions . . . 353
10.21 Auger-type intersubband processes . . . 354
10.22 Asymmetric intrasubband processes . . . 355
10.23 Empirical relationships . . . 356
10.24 A generalised expression for scattering of two-dimensional carriers . . . 357
10.25 Impurity scattering . . . 359
10.26 Alloy disorder scattering . . . 363
10.27 Alloy disorder scattering in quantum wells . . . 367
10.28 Interface roughness scattering . . . 368
10.29 Interface roughness scattering in quantum wells . . . 372
10.30 Carrier scattering in quantum wires and dots . . . 374
Exercises . . . 375
References . . . 376

11 Optical properties of quantum wells **379**
11.1 Carrier–photon scattering . . . 379
11.2 Spontaneous emission lifetime . . . 384
11.3 Intersubband absorption in quantum wells . . . 386
11.4 Bound–bound transitions . . . 389
11.5 Bound–free transitions . . . 390
11.6 Rectangular quantum well . . . 391
11.7 Intersubband optical nonlinearities . . . 394
11.8 Electric polarisation . . . 396
11.9 Intersubband second harmonic generation . . . 397
11.10 Maximisation of resonant susceptibility . . . 399
Exercises . . . 403
References . . . 403

12 Carrier transport **407**
12.1 Introduction . . . 407
12.2 Quantum cascade lasers . . . 407
12.3 Realistic quantum cascade laser . . . 413
12.4 Rate equations . . . 415
12.5 Self-consistent solution of the rate equations . . . 416
12.6 Calculation of the current density . . . 418
12.7 Phonon and carrier–carrier scattering transport . . . 419
12.8 Electron temperature . . . 420
12.9 Calculation of the gain . . . 423
12.10 QCLs, QWIPs, QDIPs and other methods . . . 425
12.11 Density matrix approaches . . . 427
12.11.1 Time evolution of the density matrix . . . 430
12.11.2 Density matrix modelling of terahertz quantum cascade lasers . . . 431
Exercises . . . 435
References . . . 437

13 Optical waveguides **441**
13.1 Introduction to optical waveguides . . . 441
13.2 Optical waveguide analysis . . . 443
13.2.1 The wave equation . . . 443
13.2.2 The transfer matrix method . . . 446
13.2.3 Guided modes in multilayer waveguides . . . 449
13.3 Optical properties of materials . . . 451
13.3.1 Semiconductors . . . 452
13.3.2 Influence of free carriers . . . 454
13.3.3 Carrier mobility model . . . 456
13.3.4 Influence of doping . . . 457
13.4 Application to waveguides of laser devices . . . 458
13.4.1 Double heterostructure laser waveguide . . . 460
13.4.2 Quantum cascade laser waveguides . . . 461

13.5 Thermal properties of waveguides . . . 466
13.6 The heat equation . . . 467
13.7 Material properties . . . 468
13.7.1 Thermal conductivity . . . 468
13.7.2 Specific heat capacity . . . 469
13.8 Finite-difference approximation to the heat equation . . . 471
13.9 Steady-state solution of the heat equation . . . 472
13.10 Time-resolved solution . . . 475
13.11 Simplified RC thermal models . . . 477
Exercises . . . 479
References . . . 481

14 Multiband envelope function (k.p) method 483
14.1 Symmetry, basis states and band structure . . . 483
14.2 Valence-band structure and the 6×6 Hamiltonian . . . 484
14.3 4×4 valence-band Hamiltonian . . . 488
14.4 Complex band structure . . . 489
14.5 Block-diagonalisation of the Hamiltonian . . . 490
14.6 The valence band in strained cubic semiconductors . . . 492
14.7 Hole subbands in heterostructures . . . 494
14.8 Valence-band offset . . . 495
14.9 The layer (transfer matrix) method . . . 496
14.10 Quantum well subbands . . . 501
14.11 The influence of strain . . . 502
14.12 Strained quantum well subbands . . . 502
14.13 Direct numerical methods . . . 503
Exercises . . . 504
References . . . 504

15 Empirical pseudo-potential band structure 505
15.1 Principles and approximations . . . 505
15.2 Elemental band structure calculation . . . 506
15.3 Spin–orbit coupling . . . 514
15.4 Compound semiconductors . . . 516
15.5 Charge densities . . . 519
15.6 Calculating the effective mass . . . 522
15.7 Alloys . . . 522
15.8 Atomic form factors . . . 524
15.9 Generalisation to a large basis . . . 525
15.10 Spin–orbit coupling within the large-basis approach . . . 528
15.11 Computational implementation . . . 529
15.12 Deducing the parameters and application . . . 530
15.13 Isoelectronic impurities in bulk . . . 533
15.14 The electronic structure around point defects . . . 539
Exercises . . . 539

References . . . 540

16 Pseudo-potential calculations of nanostructures 543
16.1 The superlattice unit cell . . . 543
16.2 Application of large-basis method to superlattices . . . 546
16.3 Comparison with envelope function approximation . . . 550
16.4 In-plane dispersion . . . 551
16.5 Interface coordination . . . 553
16.6 Strain-layered superlattices . . . 553
16.7 The superlattice as a perturbation . . . 555
16.8 Application to GaAs/AlAs superlattices . . . 560
16.9 Inclusion of remote bands . . . 562
16.10 The valence band . . . 562
16.11 Computational effort . . . 563
16.12 Superlattice dispersion and the interminiband laser . . . 564
16.13 Addition of electric field . . . 565
16.14 Application of the large-basis method to quantum wires . . . 570
16.15 Confined states . . . 572
16.16 Application of the large-basis method to tiny quantum dots . . . 573
16.17 Pyramidal quantum dots . . . 574
16.18 Transport through dot arrays . . . 575
16.19 Recent progress . . . 576
Exercises . . . 577
References . . . 578

Concluding remarks 579

A Materials parameters 581

B Introduction to the simulation tools 585
B.1 Documentation and support . . . 586
B.2 Installation and dependencies . . . 586
B.3 Simulation programs . . . 587
B.4 Introduction to scripting . . . 588
B.5 Example calculations . . . 589

Index 591

List of contributors

Principal authors

Paul Harrison developed much of his input to this text through periods at the University of Newcastle as a postgraduate research student, the University of Hull as a postdoctoral research assistant and then at the University of Leeds where he eventually became Professor of Quantum Electronics. Paul is now Pro Vice-Chancellor for Research and Innovation at Sheffield Hallam University, UK. He can always be found on the web and always answers e-mail. Currently he can be reached at:

p.harrison@shu.ac.uk or p.harrison@physics.org

Paul continues to be involved in a variety of projects, most of which centre around exploiting quantum mechanics for the creation of novel optoelectronic devices, largely, but not exclusively, in semiconductor *Quantum Wells, Wires and Dots*.

Alex Valavanis is a research fellow in the Institute of Microwaves and Photonics (IMP) at the University of Leeds, UK. He has published extensively on the theory and design of semiconductor nanostructures, and now works in the Terahertz Photonics Laboratory on the experimental development of terahertz imaging and sensor systems. His research interests include quantum cascade lasers, silicon–germanium optoelectronics, band-structure and carrier transport effects in semiconductor heterostructures, numerical methods and terahertz photonics. He can be contacted at:

a.valavanis@leeds.ac.uk

Contributing authors

Zoran Ikonić is Reader in Photonics and Plasmonics in the School of Electronic and Electrical Engineering at the University of Leeds. His research interests and experience include the full width of semiconductor physics and optoelectronic devices, in particular, band-structure calculations, strain-layered systems, carrier scattering theory, nonlinear optics, as well as conventional and quantum mechanical methods for device optimisation.

Vladimir Jovanović completed his PhD in the IMP on physical models of quantum well infrared photodetectors and quantum cascade lasers in GaN- and GaAs-based materials for near-, mid- and far-infrared (terahertz) applications.

Marco Califano is a Royal Society University Research Fellow based in the IMP at Leeds whose main interests focus on atomistic pseudo-potential modelling of the electronic and

optical properties of semiconductor nanostructures of different materials for applications in photovoltaics.

Craig A. Evans completed his PhD on the optical and thermal properties of quantum cascade lasers in the School of Electronic and Electrical Engineering, University of Leeds, in 2008. He then worked as a postdoctoral research assistant in the IMP in the field of rare-earth doped fibre lasers and integrated photonic device modelling and has now joined the staff of the school.

Dragan Indjin is Reader in Optoelectronics and Nanoscale Electronics in the School of Electronic and Electrical Engineering at the University of Leeds. He has research interests in semiconductor nanostructures, nonlinear optics, quantum computing and spintronics.

Andrew Grier completed his PhD in the School of Electronic and Electrical Engineering at the University of Leeds in 2015. His research interests include band-structure and transport calculations in quantum cascade lasers and GaN-based intersubband devices.

Preface

I have been working on theoretical and computational studies of the electronic, optical and magnetic properties of semiconductor heterostructures for two decades. During this time I have had to follow through various theoretical derivations from either books or research papers and frankly I've struggled time and time again. There never seems to be enough detail and what is commonly a couple of lines in a research paper can literally turn into many pages of precise mathematics.

With the advent of computers and their wide application to science, the emphasis of theoretical work has changed. Years ago, theoreticians aimed to produce short neat relationships, which summarised physical effects. The concentration dependency of the metal–insulator transition of Mott is an excellent example of a complex process summarised in a compact equation, which could then be evaluated by hand. The modern approach to the same problem would be to take a microscopic model of a semiconductor, calculate the band structure, and then vary the impurity concentration, thus deducing an empirical relationship from the numerics. The material system would then be changed and the process repeated. It would therefore appear that the old way was preferable, but in these times of materials modelling and device design, *quantitative results* are what we're paid for.

This book is not a book about condensed matter, nor is it a book about the physical properties of semiconductor heterostructures. It is not a book reporting the wealth of experimental measurements made on low-dimensional semiconductor systems, nor is it meant to be a general light reading book that you might cuddle up with in bed! This book isn't even a review of all the methods that are available for calculating various properties. This book is merely an in-depth look at how quantities *can* be calculated.

This book *is* aimed at providing all of the essential information, both theoretical and computational, in order that the reader can, starting from essentially nothing, understand how the electronic, optical and transport properties of semiconductor heterostructures are calculated. However, perhaps more importantly, starting from this low common denominator, this text is designed to lead the reader through a series of simple example theoretical and computational implementations, and slowly build from solid foundations, to a level where the reader can begin to initiate theoretical investigations or explanations of their own.

I believe that there are two aspects to theoretical work, with the first being to analyse and interpret experimental data, while the second is to advance new ideas. My hope is that this book will certainly facilitate the former and I believe that I will at least provide the knowledge and skills base from which quantified predictions can be developed from the beginnings of an idea.

I hope that this book will appeal to readers from outside the low-dimensional semiconductor community. Some of the examples developed will certainly be relevant to the semiconductor community at large, while the microscopic calculations presented could be of interest to other areas of condensed matter, such as carbon nanostructures, high-temperature superconductors, etc.

I have attempted here to write a book almost in the style of a mathematics course text. In such books they often describe briefly why differential equations or integration are important and then move on to show the standard techniques for solution, followed by examples and perhaps the application to real problems. Sometimes the books just state 'This is a binomial expansion and here's what you do'. In some ways this book follows both those routes. I expect that many readers will look at this present book having already a great deal of appreciation for their own particular problem. I would think that they have already quantified it in terms of, for example, 'I must calculate the exciton binding energy'. Then they'll find that this book shows them exactly how to carry this out, and indeed provides the computer codes for them to achieve their aim quickly. I wouldn't expect a reader to pick this work up knowing nothing about solid state physics—it is not aimed at that particular person—and as I've stated already, there are many excellent books available which describe in detail the electronic, optical, transport, and other properties of semiconductors and semiconductor heterostructures. It is not my aim here to compete with these in any way; indeed I hope this present work will complement the earlier works.

Therefore this book was written to fill a need, namely collecting and documenting together derivations. It is a chance to set the mathematics in stone. By incorporating all of the steps in a derivation there is no possibility of hiding away and 'glossing over' any point that is not fully understood. In many ways this means leaving oneself bare, for any mistakes or errors *will* be spotted, but this will give the opportunity for them to be corrected and hence allow the text to converge (through later editions) to a true record, which will be of use to future generations of students and researchers alike.

Realising my own limitations in the second and third editions, I invited colleagues to contribute chapters on their area of expertise. It was a pleasure to be part of a team and a thoroughly enjoyable experience. The book improved through its breadth and depth of coverage. This fourth edition has been a true partnership and has led to much new material as well as exercises to guide the reader through their development. My co-author Alex has added more depth to many of the chapters of the book, and what is not necessarily evident from the book is that he has also redeveloped most of the original computer codes, coupled with many new ones, into a professional package for semiconductor nanostructure design, and these accompaniments are availably freely on the internet.

Paul Harrison, Sheffield Hallam University

Preface to the fourth edition

I first met Paul Harrison ten years ago on a visit to the University of Leeds, when I was seeking out an interesting topic for my PhD research. I remember our enlightening discussion in The Eldon pub, in which Paul convinced me that semiconductor nanostructures are an ideal playground for anyone interested in applied mathematics or computational physics. Six months later, the first book my PhD supervisor, Rob Kelsall, recommended to me was

Quantum Wells, Wires and Dots. Since then, I have worked on a wide range of theoretical and experimental topics in the field, but this book has always remained a reliable reference guide for me.

I am greatly honoured to have been invited by Paul to join him as the author of the fourth edition. I hope my contributions to the text bring something valuable to established readers and newcomers alike. One of our main aims for this edition is to encourage the reader to explore and interact with the subject matter, to better understand the predictions and implementation of the theoretical and numerical models, and the consequences of any assumptions and approximations. As such, we have added example questions to the end of each chapter. Although some of these questions are straightforward applications or derivations of the theory, many are intended to be challenging and open-ended explorations of the subject matter. We hope that the reader will use these as starting points to explore the methods presented in this book, and indeed to find their breaking points and improve upon them!

We have also brought back the simulation tools that were provided with the first edition, but have released them online under an open-source license at `launchpad.net/qwwad`. The code has been updated and optimised, and we encourage all our readers to study it, use it in their research, modify and share it under the terms of its license. I will, of course, very happily accept any enhancements or bug fixes that are offered by readers!

In terms of new material, I have added a much more thorough narrative of band non-parabolicity effects throughout Chapters 1, 2 and 3, allowing the reader to apply our theoretical models to modern semiconductor devices that rely on high-energy electron states or narrow-bandgap materials. I have also updated Chapter 3 to reflect the relatively recent developments in robust matrix-based methods for solving the Schrödinger and Poisson equations, and have added new material to Chapter 7 to discuss the calculation of critical thickness in strained semiconductors. I have made extensive revisions to Chapter 10: it now presents a generalised model of scattering rates in two-dimensional systems, with new material describing interface roughness, impurity and alloy disorder scattering phenomena. Chapter 11 has been revised to focus more heavily on generalised models of optical properties of quantum-confined systems. My colleague, Andrew Grier, has contributed new sections to Chapter 12, which discuss density matrix methods for modelling coherent transport. Finally, I have added new sections to Chapter 13, which discuss the thermal properties of devices.

Alex Valavanis, University of Leeds

Acknowledgements

I am indebted to so many people with whom I've worked over the years that these acknowledgements have the potential for turning into a substantial work in themselves. Almost everybody with whom I've shared an office or a research project will look through this book and be able to say 'Ah yes, I helped him with that . . . '. I am truly grateful to everybody, and it is such discussions that have in many ways motivated me to write this book.

In particular, I would like to thank people such as John Davies, Winston Hagston and John Killingbeck who motivated me in my undergraduate days and spurred me on to my PhD work. At Newcastle I was lucky to be part of an excellent group with some truly great computational physicists. Perhaps the most important of these were Ian Morrison who in less than an hour really explained the bulk pseudo-potential derivation to me and Jerry Hagon who put up with my endless computing questions. I must also include my tutors, Milan Jaros and David Herbert, and, of course, my colleagues, Andy Beavis and Richard Turton.

Without doubt the largest contribution to my knowledge base came during my formative years as a postdoctoral research assistant with Winston Hagston. Besides sparking my interest in quantum mechanics as an undergraduate and then luring me back to Hull, Winston showed the way in two of the major derivations that I document here. In particular, the first of the quantum-confined impurity calculations I attempted, namely the spherically symmetric wave function, and, together with Thomas Piorek, a substantial fraction of the electron–LO phonon scattering rate.

I am indebted to all of my research colleagues at Hull, particularly the theoreticians Tom Stirner, (the already mentioned) Thomas Piorek, Richard Roberts, Fei Long, and Jabar Fatah, all for humouring and implementing my ideas for heterostructure research.

Although I had already been working with electron–phonon scattering for a number of years and had implemented this on a computer, it was my colleague Paul Kinsler who showed me where the equations that I started with actually came from. I believe this puts the phonon scattering rate work on a much surer footing and I am very grateful for his contribution.

I would also like to thank other colleagues at Leeds, in particular Bill Batty who, on comparing the results of my implementation of the shooting technique for the solution of Schrödinger's equation, actually spotted an instance when a not insignificant error arose. This forced me to look again at the iterative equation for the case of a variable effective mass and to deduce the much more stable and accurate form presented in this work. Bill Batty also deserves a mention for proofreading of my manuscript along with Kate Donovan, Marco Califano and Byron Alderman.

For the second edition I am deeply indebted to my colleagues Zoran Ikonić and Vladimir

Jovanović for contributing entirely new chapters. They are both helping to fill in gaps that this edition is certainly much better for. Some of my students have also pointed out some mistakes and opportunities for improvements: in particular, Jim McTavish, Nenad Vukmirović, Vladimir Jovanović, Ivana Savić and Craig Evans. Thanks, guys, the book is all the better for your hard work. Thanks also to all the people who contributed errata from the first edition; these are all published on the book's website.

For the third edition I am very grateful to my colleagues Marco Califano, Craig Evans and Dragan Indjin, who have contributed entirely new chapters on their specialist areas of work. In turn, Dragan Indjin would like to thank Vitomir Milanović and Jelena Radovanović for their help, and Craig Evans thanks Zoran Ikonić for many useful discussions. Again, I think the book is better for having contributions from specialists who can add more than I can on my own. They have all worked really hard to a tight deadline and I must thank them for putting up with me perpetually nagging them to finish their work. My interactions with Alex Valavanis, Leon Lever and Will Freeman, in particular, have been very valuable. Thanks, guys. Marie Barber and Zoran Ikonić deserve special mention for helping me in all aspects of my work, and without them the third edition wouldn't exist.

I would like to express my gratitude to my previous employers, the University of Leeds and the School of Electronic and Electrical Engineering, for providing me in the first instance with a University Research Fellowship, which gave me an excellent platform on which to build a career in research. It was the flexibility of this scheme that allowed me the time to undertake the first edition. Since that time my employees have continued to support me, providing me with a position in which I can earn my living through research, teaching and administration, but still have enough time left over to work on the second and third editions.

This fourth edition wouldn't have happened without the drive and enthusiasm of Alex Valavanis: I was all up for moving full-time into a leadership role, but Alex's spark lured me back for another cycle. Together we have developed the exercises that now appear at the end of each chapter, but it is Alex who has produced the new material within the chapters. He's also spruced up the diagrams, organised all the electronic files and reworked all the original computer codes into a professional open-source package. Thanks, Alex, a pleasure to work with you. Thanks also to my PhD student Razif Razali who has helped me understand nonlinear optics! Lastly, thanks to my current employers, Sheffield Hallam University, who have given me the time and space to continue with my research and scholarly pursuits.

P.H.

As a new author for the fourth edition, I owe Paul Harrison a great deal of thanks for entrusting me with this task. My thanks to Paul extend back much further, though, for introducing me to computational studies of nanostructures a decade ago, and encouraging me to begin my postgraduate studies at Leeds. I am also indebted to my PhD supervisors, Rob Kelsall and Zoran Ikonić, who supported and guided me through the first years of my research in the subject. I join Paul in thanking all my colleagues, named above, in the Institute of Microwaves and Photonics at Leeds for providing the vibrant working environment that has nurtured and supported my research.

In addition, I'd like to thank my colleagues and the students in the 'Quantum Well' who have worked with me to develop numerous computational tools for simulating heterostructures. In particular, thanks go to Jonny Cooper, Leon Lever, Viet Dinh, Craig

Evans, Andrew Grier, Pavlo Ivanov and Reeno Reeder. We've learnt a lot together on the way and the simulation tools provided with this book would not have been as complete or robust without the experience I gained from working with such a great team.

Semiconductor physics is a field in which experimental and theoretical research activities complement each other exceptionally well: there is great satisfaction to be found both in using theoretical models to underpin robust experimental activity, and in developing accurate simulations through an appreciation of practical experimental considerations. I therefore thank my experimental research colleagues in the Terahertz Photonics Laboratory, in particular Paul Dean, Josh Freeman, Andrew Burnett, Iman Kundu, Yingjun Han, Chris Wood and Lianhe Li, our principal investigators, John Cunningham, Edmund Linfield and Giles Davies, and the many talented students I have had the fortune to work with.

A.V.

Introduction

Since their discovery/invention by Esaki and Tsu in the 1970s, semiconductor quantum wells and superlattices have evolved from scientific curiosities to a means of probing the fundamentals of quantum mechanics, and more recently into wealth-creating semiconductor devices.

The aim of this book is to provide solid foundations in the theoretical methods necessary for calculating *some* of the basic electronic and optical properties of semiconductor quantum wells, wires and dots. Some background knowledge will be required; we recommend, for example, the books by Ashcroft and Mermin [1] and Blakemore [2] for support material in solid state physics, and works such as Eisberg [3] and Weidner and Sells [4] for quantum theory. The present treatise should be considered to complement existing books in the field. We also thoroughly recommend the books by Jaros [5], Davies and Long [6], Kelly [7], Turton [8], Ivchenko and Pikus [9], Shik [10] and Basu [11]. These texts will provide the literature reviews, and descriptive introductions, and also, in some cases, detailed theoretical treatments.

In this work, a brief resumé of quantum theory and solid state physics is given in Chapter 1 before launching into the main body of the book—the theoretical and computational framework of semiconductor heterostructures. Chapters 2 and 3 present methods for computing the states of carriers within heterostructures, using analytical and numerical techniques, respectively. Chapters 4, 5 and 6 then apply these methods to the study of more specialised cases: heterostructures in which the materials are allowed to diffuse between layers; systems containing charged impurities; and excitonic states, respectively. Chapter 7 discusses mechanical strain in thin semiconductor films, which has a significant influence on the behaviour of silicon–germanium and nitride-based heterostructure devices.

Chapters 8 and 9 move the focus of discussion on to quantum wire and dot systems, in which the motion of carriers is restricted to either one or zero dimensions. The following three chapters then discuss the dynamic behaviour of carriers within heterostructures, including scattering of carriers between states in Chapter 10 and their interaction with light in Chapter 11. These interactions are used in Chapter 12 to describe the current and optical gain of realistic optoelectronic devices. Chapter 13 then builds upon these models of the *internal* behaviour of devices, to discuss the optical and thermal modelling techniques used to develop complete devices within optical waveguides and heat-sinks.

The final three chapters focus on more detailed band-structure modelling techniques for computing the behaviour of carriers within a range of different energy bands and wave-vector states. Chapter 14 discusses the k.p method for extending the envelope function

approximation to multiple energy bands, and Chapter 15 describes the empirical pseudo-potential method for determining energy states based on a model of the potential around atoms in a crystal. The pseudo-potential technique is then extended to nanostructure systems in Chapter 16.

References

[1] N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Saunders College Publishing, Philadelphia, 1976.

[2] J. S. Blakemore, *Solid State Physics*, Cambridge University Press, Cambridge, Second edition, 1985.

[3] R. M. Eisberg, *Fundamentals of Modern Physics*, John Wiley & Sons, Inc., New York, 1961.

[4] R. T. Weidner and R. L. Sells, *Elementary Modern Physics*, Allyn and Bacon, Boston, Third edition, 1980.

[5] M. Jaros, *Physics and Applications of Semiconductor Microstructures*, Clarendon Press, Oxford, 1989.

[6] J. H. Davies and A. R. Long, Eds., *Physics of Nanostructures*, IOP Publishing, Bristol, 1992.

[7] M. J. Kelly, *Low Dimensional Semiconductors: Materials, Physics, Technology, Devices*, Clarendon Press, Oxford, 1995.

[8] R. Turton, *The Quantum Dot: a Journey into the Future of Microelectronics*, Oxford University Press, New York, 1995.

[9] E. L. Ivchenko and G. Pikus, *Superlattices and Other Heterostructures: Symmetry and Optical Phenomena*, Springer-Verlag, Berlin, Second edition, 1995.

[10] A. Shik, *Quantum Wells: Physics and Electronics of Two-Dimensional Systems*, World Scientific, London, 1997.

[11] P. K. Basu, *Theory of Optical Processes in Semiconductors*, Clarendon Press, Oxford, 1997.

1

Semiconductors and heterostructures

1.1 The mechanics of waves

De Broglie (see reference [1]) stated that a particle of momentum p has an associated wave of wavelength λ given by:

$$\lambda = \frac{h}{p} \tag{1.1}$$

Thus, an electron in a vacuum at a position $\mathbf{r}$ and away from the influence of any electromagnetic potentials could be described by a *state function*, which is of the form of a wave, i.e.

$$\psi = \mathrm{e}^{i(\mathbf{k}_\bullet \mathbf{r} - \omega t)} \tag{1.2}$$

where t is the time, ω the angular frequency and the modulus of the wave vector is given by:

$$k = |\mathbf{k}| = \frac{2\pi}{\lambda} \tag{1.3}$$

The quantum mechanical momentum has been deduced to be a linear operator [2] acting upon the *wave function* ψ, with the momentum $\mathbf{p}$ arising as an eigenvalue, i.e.

$$-i\hbar\nabla\psi = \mathbf{p}\psi \tag{1.4}$$

where

$$\nabla = \frac{\partial}{\partial x}\hat{\mathbf{i}} + \frac{\partial}{\partial y}\hat{\mathbf{j}} + \frac{\partial}{\partial z}\hat{\mathbf{k}} \tag{1.5}$$

which, when operating on the electron vacuum wave function in equation (1.2), would give the following:

$$-i\hbar\nabla \mathrm{e}^{i(\mathbf{k}_\bullet \mathbf{r} - \omega t)} = \mathbf{p}\mathrm{e}^{i(\mathbf{k}_\bullet \mathbf{r} - \omega t)} \tag{1.6}$$

and therefore

$$-i\hbar\left(\frac{\partial}{\partial x}\hat{\mathbf{i}} + \frac{\partial}{\partial y}\hat{\mathbf{j}} + \frac{\partial}{\partial z}\hat{\mathbf{k}}\right)\mathrm{e}^{i(k_x x + k_y y + k_z z - \omega t)} = \mathbf{p}\mathrm{e}^{i(\mathbf{k}_\bullet \mathbf{r} - \omega t)} \tag{1.7}$$

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

$$\therefore -i\hbar\left(ik_x\hat{\mathbf{i}} + ik_y\hat{\mathbf{j}} + ik_z\hat{\mathbf{k}}\right)\mathrm{e}^{i(k_x x+k_y y+k_z z-\omega t)} = \mathbf{p}\mathrm{e}^{i(\mathbf{k}\bullet\mathbf{r}-\omega t)} \tag{1.8}$$

Thus the eigenvalue

$$\mathbf{p} = \hbar\left(k_x\hat{\mathbf{i}} + k_y\hat{\mathbf{j}} + k_z\hat{\mathbf{k}}\right) = \hbar\mathbf{k} \tag{1.9}$$

which, not surprisingly, can be simply manipulated ($p = \hbar k = (h/2\pi)(2\pi/\lambda)$) to reproduce de Broglie's relationship in equation (1.1).

Following on from this, classical mechanics gives the kinetic energy of a particle of mass m as:

$$T = \frac{1}{2}mv^2 = \frac{(mv)^2}{2m} = \frac{p^2}{2m} \tag{1.10}$$

Therefore it may be expected that the quantum mechanical analogy can also be represented by an eigenvalue equation with an operator:

$$\frac{1}{2m}\left(-i\hbar\nabla\right)^2\psi = T\psi \tag{1.11}$$

i.e.

$$-\frac{\hbar^2}{2m}\nabla^2\psi = T\psi \tag{1.12}$$

where T is the kinetic energy eigenvalue, and, given the form of ∇ in equation (1.5), then:

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{1.13}$$

When acting upon the electron vacuum wave function, i.e.

$$-\frac{\hbar^2}{2m}\nabla^2\mathrm{e}^{i(\mathbf{k}\bullet\mathbf{r}-\omega t)} = T\mathrm{e}^{i(\mathbf{k}\bullet\mathbf{r}-\omega t)} \tag{1.14}$$

then

$$-\frac{\hbar^2}{2m}\left(i^2k_x^2 + i^2k_y^2 + i^2k_z^2\right)\mathrm{e}^{i(\mathbf{k}\bullet\mathbf{r}-\omega t)} = T\mathrm{e}^{i(\mathbf{k}\bullet\mathbf{r}-\omega t)} \tag{1.15}$$

Thus the kinetic energy eigenvalue is given by:

$$T = \frac{\hbar^2k^2}{2m} \tag{1.16}$$

For an electron in a vacuum away from the influence of electromagnetic fields, the total energy E is just the kinetic energy T. Thus the dispersion or energy versus momentum (which is proportional to the wave vector k) curves are parabolic, just as for classical free particles, as illustrated in Fig. 1.1.

The equation describing the total energy of a particle in this wave description is called the time-independent Schrödinger equation and, for this case with only a kinetic energy contribution, can be summarised as follows:

$$-\frac{\hbar^2}{2m}\nabla^2\psi = E\psi \tag{1.17}$$

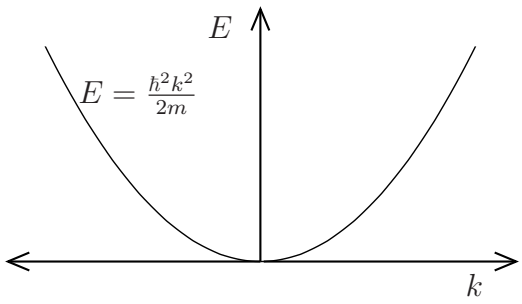


Figure 1.1: The energy versus wave vector (proportional to momentum) curve for an electron in a vacuum

A corresponding equation also exists that includes the time dependency explicitly; this is obtained by operating on the wave function by the linear operator $i\hbar\partial/\partial t$, i.e.

$$i\hbar\frac{\partial}{\partial t}\mathrm{e}^{i(\mathbf{k}\bullet\mathbf{r}-\omega t)} = i\hbar(-i\omega)\mathrm{e}^{i(\mathbf{k}\bullet\mathbf{r}-\omega t)} \tag{1.18}$$

or

$$i\hbar\frac{\partial}{\partial t}\psi = \hbar\omega\psi \tag{1.19}$$

Clearly, this eigenvalue $\hbar\omega$ is also the total energy but in a form usually associated with waves, e.g. a photon. These two operations on the wave function represent the two complementary descriptions associated with *wave–particle duality*. Thus the second, i.e. time-dependent, Schrödinger equation is given by:

$$i\hbar\frac{\partial}{\partial t}\psi = E\psi \tag{1.20}$$

1.2 Crystal structure

The vast majority of the mainstream semiconductors have a face-centred cubic Bravais lattice, as illustrated in Fig 1.2. The lattice points are defined in terms of linear combinations of a set of *primitive lattice vectors*, one choice for which is:

$$\mathbf{a}_1 = \frac{A_0}{2}(\hat{\mathbf{j}} + \hat{\mathbf{k}}), \qquad \mathbf{a}_2 = \frac{A_0}{2}(\hat{\mathbf{k}} + \hat{\mathbf{i}}), \qquad \mathbf{a}_3 = \frac{A_0}{2}(\hat{\mathbf{i}} + \hat{\mathbf{j}}) \tag{1.21}$$

The *lattice vectors* then follow as the set of vectors:

$$\mathbf{R} = \alpha_1\mathbf{a}_1 + \alpha_2\mathbf{a}_2 + \alpha_3\mathbf{a}_3 \tag{1.22}$$

where α_1, α_2, and α_3 are integers.

The complete crystal structure is obtained by placing the atomic basis at each Bravais lattice point. For materials such as Si, Ge, GaAs, AlAs and InP, this consists of two atoms, one at $(\frac{1}{8},\frac{1}{8},\frac{1}{8})$ and the other at $(-\frac{1}{8},-\frac{1}{8},-\frac{1}{8})$, in units of A_0.

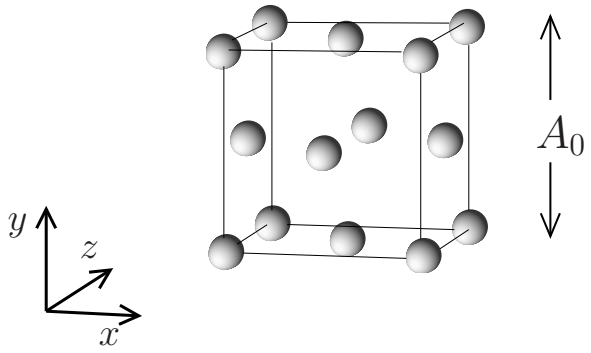


Figure 1.2: The face-centred cubic Bravais lattice

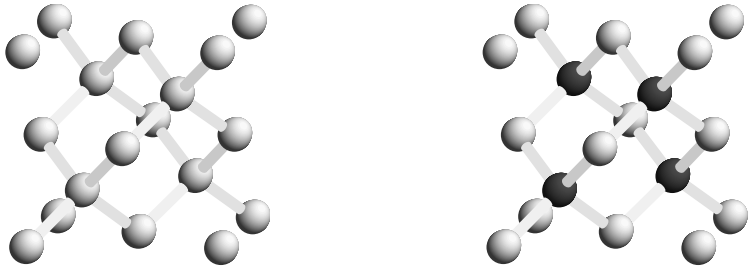


Figure 1.3: The diamond (left) and zinc blende (right) crystal structures

For the group IV materials, such as Si and Ge, as the atoms within the basis are the same, the crystal structure is equivalent to diamond (see Fig. 1.3 (left)). For III–V and II–VI compound semiconductors such as GaAs, AlAs, InP, HgTe and CdTe, the cation sits on the $(-\frac{1}{8},-\frac{1}{8},-\frac{1}{8})$ site and the anion on $(+\frac{1}{8},+\frac{1}{8},+\frac{1}{8})$; this type of crystal is called the *zinc blende* structure, after ZnS (see Fig. 1.3 (right)). The only exception to this rule is GaN, and its important $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{N}$ alloys, which have risen to prominence in recent years due to their use in green and blue light emitting diodes and lasers (see, for example, [3]); these materials have the *wurtzite* structure (see [4], p. 47).

From an electrostatics viewpoint, the crystal potential consists of a three-dimensional lattice of spherically symmetric ionic core potentials screened by the inner shell electrons (see Fig. 1.4), which are further surrounded by the covalent bond charge distributions that hold everything together.

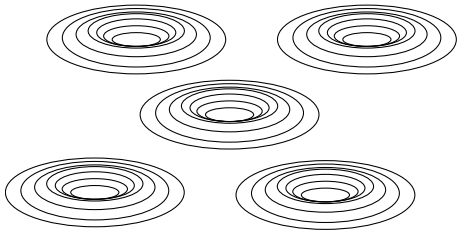


Figure 1.4: Schematic illustration of the ionic core component of the crystal potential across the $\{001\}$ planes—a three-dimensional array of spherically symmetric potentials

1.3 The effective mass approximation

Therefore, the crystal potential is complicated; however, using the principle of simplicity,[1] imagine that it can be approximated by a constant! Then the Schrödinger equation derived for an electron in a vacuum would be applicable. Clearly, though, a crystal is not a vacuum so allow the introduction of an empirical fitting parameter called the effective mass, m^*. Thus the time-independent Schrödinger equation becomes:

$$-\frac{\hbar^2}{2m^*}\nabla^2\psi = E\psi \tag{1.23}$$

and the energy solutions follow as:

$$E = \frac{\hbar^2 k^2}{2m^*} \tag{1.24}$$

This is known as the *effective mass approximation* and has been found to be very suitable for relatively low electron momenta as occur with low electric fields. Indeed, it is the most widely used parameterisation in semiconductor physics (see any good solid state physics book, e.g. [4, 5, 6]). Experimental measurements of the effective mass have revealed it to be anisotropic—as might be expected since the crystal potential along, say, the [001] axis is different than along the [111] axis. Adachi [7] collates reported values for GaAs and its alloys; the effective mass in other materials can be found in Landolt and Börnstein [8].

In GaAs, the reported effective mass is around 0.067 m_0, where m_0 is the rest mass of an electron. Figure 1.5 plots the dispersion curve for this effective mass, in comparison with that of an electron in a vacuum.

1.4 Band theory

It has also been found from experiment that there are two distinct energy *bands* within semiconductors. The lower band is almost full of electrons and can conduct by the movement of the empty states. This band originates from the valence electron states which constitute the covalent bonds holding the atoms together in the crystal. In many ways, electric charge in a solid resembles a fluid, and the analogy for this band, labelled the *valence band*, is that the empty states behave like bubbles within the fluid—hence their name, *holes*.

[1] Choose the simplest thing first; if it works, use it, and if it doesn't, then try the next simplest!

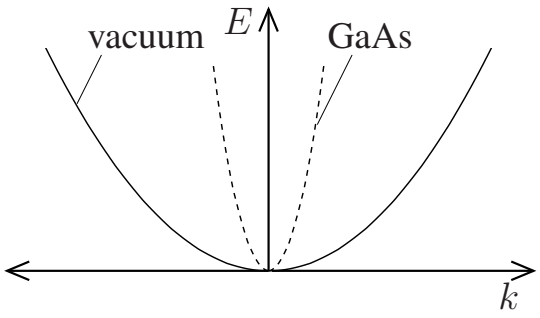


Figure 1.5: The energy versus wave vector (proportional to momentum) curves for an electron in GaAs compared to that in a vacuum

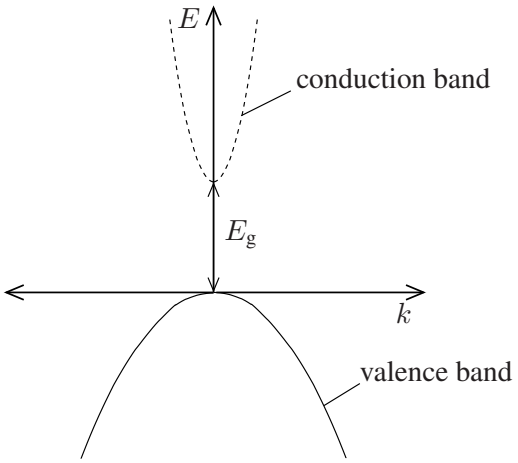


Figure 1.6: The energy versus wave vector curves for an electron in the conduction band and a hole in the valence band of GaAs

In particular, the holes rise to the uppermost point of the valence band, and just as it is possible to consider the release of carbon dioxide through the motion of beer in a glass, it is actually easier to study the motion of the bubble (the absence of beer), or in this case the motion of the hole.

In a semiconductor, the upper band is almost devoid of electrons. It represents excited electron states which are occupied by electrons promoted from localised covalent bonds into extended states in the body of the crystal. Such electrons are readily accelerated by an applied electric field and contribute to current flow. This band is therefore known as the *conduction band*.

Figure 1.6 illustrates these two bands. Notice how the valence band is inverted—this is a reflection of the fact that the 'bubbles' rise to the top, i.e. their lowest-energy states are at

the top of the band. The energy difference between the two bands is known as the *band gap*, labelled as E_{gap} on the figure. The particular curvatures used in both bands are indicative of those measured experimentally for GaAs, namely effective masses of around 0.067 m_0 for an electron in the conduction band, and 0.6 m_0 for a (heavy) hole in the valence band. The convention is to put the zero of the energy at the top of the valence band. Note the extra qualifier 'heavy'. In fact, there is more than one valence band, and they are distinguished by their different effective masses. Chapter 15 will discuss band structure in more detail; this will be in the context of a microscopic model of the crystal potential which goes beyond the simple ideas introduced here.

1.5 Heterojunctions

The effective mass approximation is for a bulk crystal, which means the crystal is so large with respect to the scale of an electron wave function that it is effectively infinite. Within the effective mass approximation, the Schrödinger equation has been found to be as follows:

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial z^2}\psi(z) = E\psi(z) \tag{1.25}$$

When two such materials are placed adjacent to each other to form a *heterojunction*, this equation is valid within each, remembering of course that the effective mass could be a function of position. However, the band gaps of the materials can also be different (see Fig. 1.7).

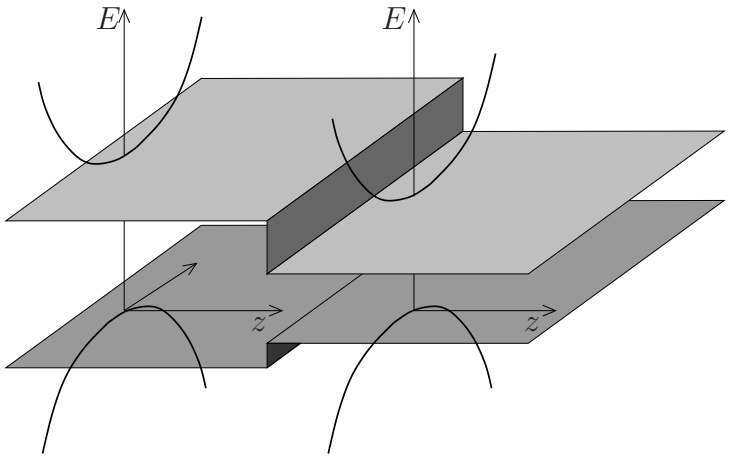


Figure 1.7: Two dissimilar semiconductors with different band gaps joined to form a heterojunction; the curves represent the unrestricted motion parallel to the interface

The discontinuity in either the conduction or the valence band can be represented by a constant potential term. Thus the Schrödinger equation for any one of the bands, taking the

effective mass to be the same in each material, would be generalised to:

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial z^2}\psi(z) + V(z)\psi(z) = E\psi(z) \tag{1.26}$$

In the above example, the one-dimensional potentials $V(z)$ representing the band discontinuities at the heterojunction would have the form shown in Fig. 1.8, noting that increasing hole energy in the valence band is measured downwards.

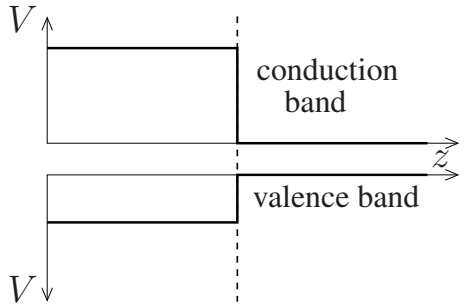


Figure 1.8: The one-dimensional potentials $V(z)$ in the conduction and valence band as might occur at a heterojunction (marked with a dashed line) between two dissimilar materials

1.6 Heterostructures

Heterostructures are formed from multiple heterojunctions, and thus a myriad of possibilities exist. If a thin layer of a narrower-bandgap material (A, say) is sandwiched between two layers of a wider-bandgap material (B), as illustrated in Fig. 1.9 (left), then they form a double heterojunction. If layer A is sufficiently thin for *quantum properties* to be exhibited, then such a band alignment is called a *single quantum well*.

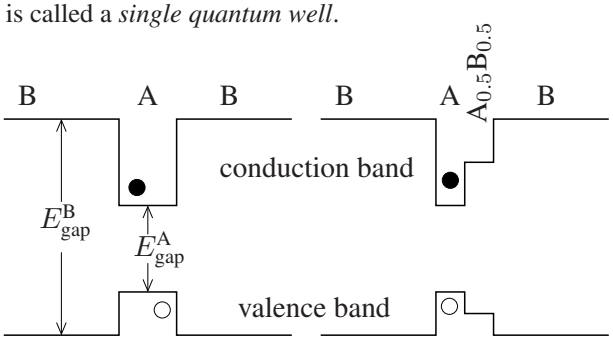


Figure 1.9: The one-dimensional potentials $V(z)$ in the conduction and valence bands for a typical single quantum well (left) and a stepped quantum well (right)

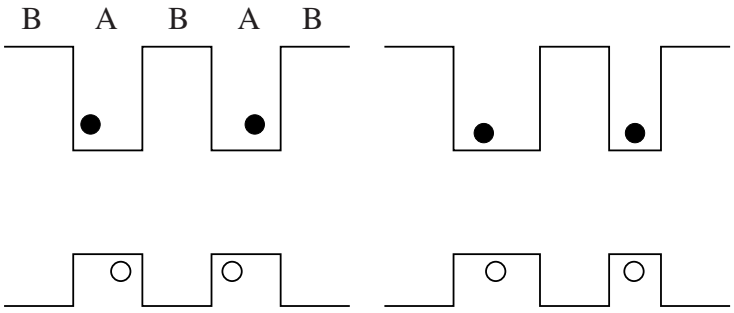


Figure 1.10: The one-dimensional potentials $V(z)$ in the conduction and valence band for typical *symmetric* (left) and *asymmetric* (right) double quantum wells

If any charge carriers exist in the system, whether thermally produced intrinsic or extrinsic as the result of doping, they will attempt to lower their energies. Hence in this example, any electrons (solid circles) or holes (open circles) will collect in the quantum well (see Fig. 1.9). Additional semiconductor layers can be included in the heterostructure, for example a *stepped* or *asymmetric* quantum well can be formed by the inclusion of an alloy between materials A and B, as shown in Fig. 1.9 (right).

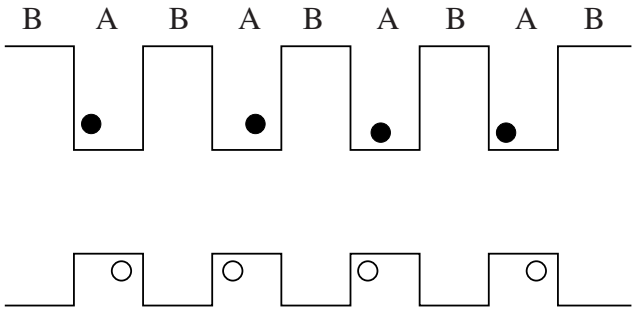


Figure 1.11: The one-dimensional potentials $V(z)$ in the conduction and valence band for a typical multiple quantum well or superlattice

Still more complex structures can be formed, such as symmetric or asymmetric *double quantum wells*, (see Fig. 1.10) and *multiple quantum wells* or *superlattices* (see Fig. 1.11). The difference between the latter is the extent of the interaction between the quantum wells; in particular, a multiple quantum well exhibits the properties of a collection of isolated single quantum wells, whereas in a superlattice the quantum wells do interact. The motivation behind introducing increasingly complicated structures is an attempt to tailor the electronic and optical properties of these materials for exploitation in devices. Perhaps the most complicated layer structure to date is the *chirped superlattice* active region of a mid-infrared laser [9].

All of the structures illustrated so far have been examples of Type-I systems. In this type, the band gap of one material is nestled entirely within that of the wider-bandgap material. The consequence of this is that any electrons or holes fall into quantum wells which are *within* the same layer of material. Thus both types of charge carrier are localised in the same region of space, which makes for efficient (fast) recombination. However, other possibilities can exist, as illustrated in Fig. 1.12.

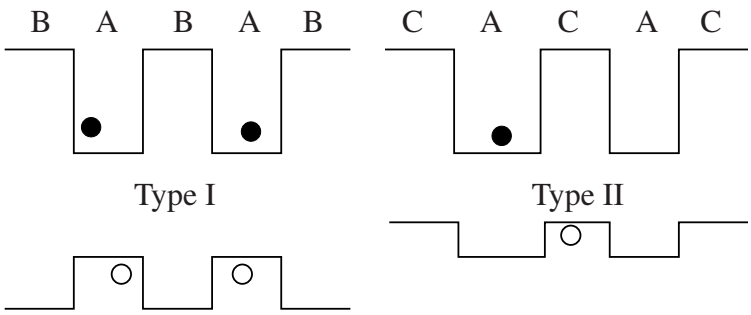


Figure 1.12: The one-dimensional potentials $V(z)$ in the conduction and valence bands for a typical Type-I superlattice (left) compared to a Type-II system (right)

In Type-II systems the band gaps of the materials (say, A and C) are aligned such that the quantum wells formed in the conduction and valence bands are in different materials, as illustrated in Fig. 1.12 (right). This leads to the electrons and holes being confined in *different* layers of the semiconductor. The consequence of this is that the recombination times of electrons and holes are long.

1.7 The envelope function approximation

Two important points have been argued:

1. The effective mass approximation *is* a valid description of bulk materials.
2. Heterojunctions between dissimilar materials, both of which can be well represented by the effective mass approximation, can be described by a material potential which derives from the difference in the band gaps.

The logical extension to point 2 is that the crystal potential of multiple heterojunctions can also be described in this manner, as illustrated extensively in the previous section.

Once this is accepted, then the electronic structure *can* be represented by the simple one-dimensional Schrödinger equation that has been aspired to:

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial z^2}\psi(z) + V(z)\psi(z) = E\psi(z) \tag{1.27}$$

The *envelope function approximation* is the name given to the mathematical justification for this series of arguments (see, for example, works by Bastard [10, 11] and Burt [12, 13]).

The name derives from the deduction that physical properties can be derived from the slowly varying envelope function, identified here as $\psi(z)$, rather than the total wave function $\psi(z)u(z)$, where the latter is rapidly varying on the scale of the crystal lattice. The validity of the envelope function approximation is still an active area of research [13]. With the line of reasoning used here, it is clear that the envelope function approximation *can* be thought of as an approximation on the material and *not* the quantum mechanics.

Some thought is enough to appreciate that the envelope function approximation will have limitations, and that these will occur for very thin layers of material. The materials *are* made of a collection of a large number of atomic potentials, so when a layer becomes thin, these individual potentials will become significant and the global average of representing the crystal potential by a constant will breakdown (see, for example, [14]). However, for the majority of examples this approach works well; this will be demonstrated in later chapters, and, in particular, a detailed comparison with an alternative approach which does account fully for the microscopic crystal potential will be made in Chapter 16.

1.8 Band non-parabolicity

For semiconductor heterostructures with relatively low barrier heights and low carrier densities, the carriers cluster at energies near the conduction-band edge, i.e. within a couple of hundred millielectronvolts, compared to a band gap of the order of $1.5\,\mathrm{eV}$. In this region, the band edge can be described by a parabolic E–$\mathbf{k}$ curve, i.e. in the form given by equation (1.24).

However, in situations where the carriers are forced up to higher energies, either by large barrier heights and narrow wells (as discussed in Chapter 2), or by very high carrier densities or high temperatures, equation (1.24) is only an approximate description of the carrier dispersion. This is especially important for holes in the valence band. The approximation can be improved upon by adding summation terms into the polynomial expansion for the energy. For example, no matter how complex the band structure along a particular direction, it clearly has inversion symmetry and hence can always be represented by an expansion in even powers of k, if sufficient terms are included, i.e. the energy E can be described by:

$$E = a_0k^0 + a_2k^2 + a_4k^4 + a_6k^6 + a_8k^8 + \ldots = \sum_{i=0}^{\infty} a_{2i}k^{2i} \tag{1.28}$$

Usually when discussing single band models, as has been focused on entirely so far, the energy origin is set at the bottom of the band of interest, and thus $a_0 = 0$. In addition, truncating the series at k^2 gives $a_2 = \hbar^2/(2m^*)$.

The next best approximation is to include terms in k^4, hence accounting for band non-parabolicity, as displayed in Fig. 1.13, with:

$$E = a_2k^2 + a_4k^4 = \frac{\hbar^2}{2m^*}\left(k^2 + \beta k^4\right) \tag{1.29}$$

where

$$\beta = a_2\frac{2m^*}{\hbar^2} \tag{1.30}$$

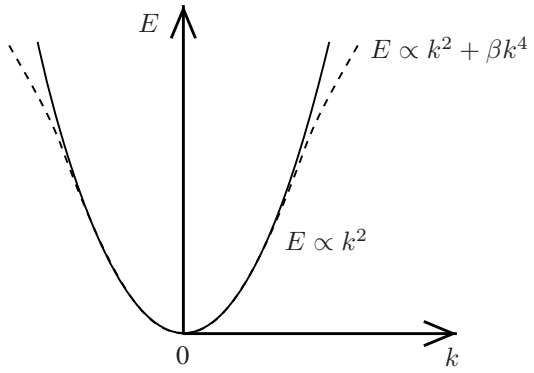


Figure 1.13: Dispersion relations for parabolic (solid line) and non-parabolic band (dashed line) models

Using the basic definition of effective mass, i.e.

$$m^* = \hbar^2 \left(\frac{\partial^2 E}{\partial k^2} \right)^{-1} \tag{1.31}$$

then clearly m^* remains a function of k, unlike the case of parabolic bands where it is a constant. As it is a function of k, then it is also a function of energy, and indeed band non-parabolicity is accounted for by an energy-dependent effective mass [15]:

$$m^*(E) = m^*(0)[1 + \alpha(E - V)] \tag{1.32}$$

where V is the band-edge potential.

The non-parabolicity parameter, α, is usually greatest when the band edge under consideration is close to other energy bands in the material. Precise values of α may be obtained experimentally or extracted from more detailed models of the band structure, such as those considered in Chapters 14 and 16. However, simple approximations exist near the edge of the conduction band, such as that given by [16, 17]:

$$\alpha = \left[1 - \frac{m^*(0)}{m_0} \right]^2 \Big/ E_{\mathrm{g}} \tag{1.33}$$

where E_{g} is the semiconductor band gap, or the even simpler model given by [18, 19]:

$$\alpha = 1/E_{\mathrm{g}} \tag{1.34}$$

The latter expression can be used to estimate the significance of non-parabolicity in a range of different materials. In GaAs, for example, $E_{\mathrm{g}} = 1.426\,\mathrm{eV}$, giving $\alpha = 0.701\,\mathrm{eV}^{-1}$. The effective mass of an electron $100\,\mathrm{meV}$ above the conduction-band edge in GaAs, therefore, has an effective mass 7% higher than that at the band edge. By comparison, in InSb, $E_{\mathrm{g}} = 0.18\,\mathrm{eV}$ and $\alpha = 5.6\,\mathrm{eV}^{-1}$, meaning that an equivalent electron has an effective mass 56% higher than that at the band edge. This illustrates that although non-parabolicity may be

very significant in narrow-bandgap semiconductors, it can be ignored safely for low-energy electrons in many common materials.

For this reason, non-parabolicity is neglected throughout most of this book, and the focus is placed on the much more elegant and immediately intuitive models of the behaviour of carriers with parabolic dispersion. However, Chapter 2 illustrates how non-parabolicity affects analytical models of certain phenomena, and Chapter 3 describes how it may be included in numerical methods.

1.9 The reciprocal lattice

For later discussions the concept of the *reciprocal lattice* needs to be developed. It has already been shown that considering electron wave functions as plane waves ($e^{i\mathbf{k}\bullet\mathbf{r}}$), as found in a vacuum, but with a correction factor called the effective mass, is a useful method of approximating the electronic band structure. In general, such a wave will not have the periodicity of the crystal lattice; however, for certain wave vectors it will. Such a set of wave vectors $\mathbf{G}$ are known as the *reciprocal lattice vectors* with the set of points mapped out by these primitives known as the *reciprocal lattice*.

If the set of vectors $\mathbf{G}$ *did* have the periodicity of the lattice, then this would imply that:

$$e^{i\mathbf{G}\bullet(\mathbf{r}+\mathbf{R})} = e^{i\mathbf{G}\bullet\mathbf{r}} \tag{1.35}$$

i.e. an electron with this wave vector $\mathbf{G}$ would have a wave function equal at all points in real space separated by a Bravais lattice vector $\mathbf{R}$. Therefore:

$$e^{i\mathbf{G}\bullet\mathbf{r}}e^{i\mathbf{G}\bullet\mathbf{R}} = e^{i\mathbf{G}\bullet\mathbf{r}} \tag{1.36}$$

which implies that

$$\mathbf{G}\bullet\mathbf{R} = 2\pi n, \qquad n \in \mathbb{Z} \tag{1.37}$$

Now learning from the form for the Bravais lattice vectors $\mathbf{R}$ given earlier in equation (1.22), it might be expected that the reciprocal lattice vectors $\mathbf{G}$ could be constructed in a similar manner from a set of three *primitive reciprocal lattice vectors*, i.e.

$$\mathbf{G} = \beta_1\mathbf{b}_1 + \beta_2\mathbf{b}_2 + \beta_3\mathbf{b}_3, \qquad \beta_1, \beta_2, \beta_3 \in \mathbb{Z} \tag{1.38}$$

With these choices, then, the primitive reciprocal lattice vectors can be written as follows:

$$\mathbf{b}_1 = 2\pi\frac{\mathbf{a}_2 \times \mathbf{a}_3}{\mathbf{a}_1\bullet(\mathbf{a}_2 \times \mathbf{a}_3)} \tag{1.39}$$

$$\mathbf{b}_2 = 2\pi\frac{\mathbf{a}_3 \times \mathbf{a}_1}{\mathbf{a}_1\bullet(\mathbf{a}_2 \times \mathbf{a}_3)} \tag{1.40}$$

$$\mathbf{b}_3 = 2\pi\frac{\mathbf{a}_1 \times \mathbf{a}_2}{\mathbf{a}_1\bullet(\mathbf{a}_2 \times \mathbf{a}_3)} \tag{1.41}$$

It is possible to verify that these forms do satisfy equation (1.37):

$$\mathbf{G}\bullet\mathbf{R} = (\beta_1\mathbf{b}_1 + \beta_2\mathbf{b}_2 + \beta_3\mathbf{b}_3)\bullet(\alpha_1\mathbf{a}_1 + \alpha_2\mathbf{a}_2 + \alpha_3\mathbf{a}_3) \tag{1.42}$$

Now $\mathbf{b}_1$ is perpendicular to both $\mathbf{a}_2$ and $\mathbf{a}_3$, and so only the product of $\mathbf{b}_1$ with $\mathbf{a}_1$ is non-zero, and similarly for $\mathbf{b}_2$ and $\mathbf{b}_3$. Hence:

$$\mathbf{G}_{\bullet}\mathbf{R} = \beta_1\alpha_1\mathbf{b}_{1\bullet}\mathbf{a}_1 + \beta_2\alpha_2\mathbf{b}_{2\bullet}\mathbf{a}_2 + \beta_3\alpha_3\mathbf{b}_{3\bullet}\mathbf{a}_3 \tag{1.43}$$

and in fact, the products $\mathbf{b}_{\mathrm{i}\bullet}\mathbf{a}_{\mathrm{i}} = 2\pi$. Therefore:

$$\mathbf{G}_{\bullet}\mathbf{R} = 2\pi\left(\beta_1\alpha_1 + \beta_2\alpha_2 + \beta_3\alpha_3\right) \tag{1.44}$$

Clearly $\beta_1\alpha_1 + \beta_2\alpha_2 + \beta_3\alpha_3$ is an integer, and hence equation (1.37) is satisfied.

Using the face-centred cubic lattice vectors defined in equation (1.21), then:

$$\mathbf{a}_{1\bullet}(\mathbf{a}_2 \times \mathbf{a}_3) = \begin{vmatrix} a_{1x} & a_{1y} & a_{1z} \\ a_{2x} & a_{2y} & a_{2z} \\ a_{3x} & a_{3y} & a_{3z} \end{vmatrix} = \begin{vmatrix} 0 & \frac{A_0}{2} & \frac{A_0}{2} \\ \frac{A_0}{2} & 0 & \frac{A_0}{2} \\ \frac{A_0}{2} & \frac{A_0}{2} & 0 \end{vmatrix} \tag{1.45}$$

which gives:

$$\mathbf{a}_{1\bullet}(\mathbf{a}_2 \times \mathbf{a}_3) = 0 \times \begin{vmatrix} 0 & \frac{A_0}{2} \\ \frac{A_0}{2} & 0 \end{vmatrix} - \frac{A_0}{2}\begin{vmatrix} \frac{A_0}{2} & \frac{A_0}{2} \\ \frac{A_0}{2} & 0 \end{vmatrix} + \frac{A_0}{2}\begin{vmatrix} \frac{A_0}{2} & 0 \\ \frac{A_0}{2} & \frac{A_0}{2} \end{vmatrix} \tag{1.46}$$

$$\therefore \mathbf{a}_{1\bullet}(\mathbf{a}_2 \times \mathbf{a}_3) = 0 - \frac{A_0}{2}\left[0 - \left(\frac{A_0}{2}\right)^2\right] + \frac{A_0}{2}\left[\left(\frac{A_0}{2}\right)^2 - 0\right] \tag{1.47}$$

$$\therefore \mathbf{a}_{1\bullet}(\mathbf{a}_2 \times \mathbf{a}_3) = 2\left(\frac{A_0}{2}\right)^3 \tag{1.48}$$

Therefore, the first of the primitive reciprocal lattice vectors follows as:

$$\mathbf{b}_1 = 2\pi\left[2\left(\frac{A_0}{2}\right)^3\right]^{-1}\begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ \frac{A_0}{2} & 0 & \frac{A_0}{2} \\ \frac{A_0}{2} & \frac{A_0}{2} & 0 \end{vmatrix} \tag{1.49}$$

$$\begin{aligned} \therefore \mathbf{b}_1 &= 2\pi\frac{1}{2}\left(\frac{2}{A_0}\right)^3 \\ &\times\left\{\left[0 - \left(\frac{A_0}{2}\right)^2\right]\hat{\mathbf{i}} - \left[0 - \left(\frac{A_0}{2}\right)^2\right]\hat{\mathbf{j}} + \left[\left(\frac{A_0}{2}\right)^2 - 0\right]\hat{\mathbf{k}}\right\} \end{aligned} \tag{1.50}$$

$$\therefore \mathbf{b}_1 = \frac{2\pi}{A_0}\left(-\hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}}\right) \tag{1.51}$$

A similar calculation of the remaining primitive reciprocal lattice vectors $\mathbf{b}_2$ and $\mathbf{b}_3$ gives the complete set as follows:

$$\mathbf{b}_1 = \frac{2\pi}{A_0}(-\hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}}), \qquad \mathbf{b}_2 = \frac{2\pi}{A_0}(\hat{\mathbf{i}} - \hat{\mathbf{j}} + \hat{\mathbf{k}}), \qquad \mathbf{b}_3 = \frac{2\pi}{A_0}(\hat{\mathbf{i}} + \hat{\mathbf{j}} - \hat{\mathbf{k}}) \tag{1.52}$$

which are, of course, equivalent to the body-centred cubic Bravais lattice vectors (see [5], p. 68). Thus the reciprocal lattice constructed from the linear combinations:

$$\mathbf{G} = \beta_1\mathbf{b}_1 + \beta_2\mathbf{b}_2 + \beta_3\mathbf{b}_3, \qquad \beta_1, \beta_2, \beta_3 \in \mathbb{Z} \tag{1.53}$$

Table 1.1 Generation of the reciprocal lattice vectors for the face-centred cubic crystal by the systematic selection of the integer coefficients β_1, β_2 and β_3

β_1	β_2	β_3	Reciprocal lattice vector, $\mathbf{G}$ $(2\pi/A_0)$
0	0	0	$(0,0,0)$
1	0	0	$(\bar{1},1,1)$
0	1	0	$(1,\bar{1},1)$
0	0	1	$(1,1,\bar{1})$
-1	0	0	$(1,\bar{1},\bar{1})$
0	-1	0	$(\bar{1},1,\bar{1})$
0	0	-1	$(\bar{1},\bar{1},1)$
1	1	1	$(1,1,1)$
-1	-1	-1	$(\bar{1},\bar{1},\bar{1})$
1	1	0	$(0,0,2)$
1	0	1	$(0,2,0)$
0	1	1	$(2,0,0)$
-1	1	0	$(2,\bar{2},2)$
(etc.)			

is a body-centred cubic lattice with lattice constant $4\pi/A_0$.

Taking the face-centred cubic primitive reciprocal lattice vectors in equation (1.52), then:

$$\mathbf{G} = \frac{2\pi}{A_0}\left[\beta_1(-\hat{\mathbf{i}}+\hat{\mathbf{j}}+\hat{\mathbf{k}}) + \beta_2(\hat{\mathbf{i}}-\hat{\mathbf{j}}+\hat{\mathbf{k}}) + \beta_3(\hat{\mathbf{i}}+\hat{\mathbf{j}}-\hat{\mathbf{k}})\right] \tag{1.54}$$

$$\therefore \mathbf{G} = \frac{2\pi}{A_0}\left[(-\beta_1+\beta_2+\beta_3)\hat{\mathbf{i}} + (\beta_1-\beta_2+\beta_3)\hat{\mathbf{j}} + (\beta_1+\beta_2-\beta_3)\hat{\mathbf{k}}\right] \tag{1.55}$$

The specific reciprocal lattice vectors are therefore generated by taking different combinations of the integers β_1, β_2, and β_3. This is illustrated in Table 1.1.

It was shown by von Laue that, when waves in a periodic structure satisfied the following:

$$\mathbf{k}_\bullet \hat{\mathbf{G}} = \frac{1}{2}|\mathbf{G}| \tag{1.56}$$

diffraction would occur (see [5], p. 99). Thus the 'free' electron dispersion curves of earlier (Fig. 1.5), will be perturbed when the electron wave vector satisfies equation (1.56). Along the [001] direction, the smallest reciprocal lattice vector $\mathbf{G}$ is (0,0,2) (in units of $2\pi/A_0$). Substituting into equation (1.56) gives:

$$\mathbf{k}_\bullet(0\hat{\mathbf{i}}+0\hat{\mathbf{j}}+\hat{\mathbf{k}}) = \frac{1}{2}\times 2\times\frac{2\pi}{A_0} \tag{1.57}$$

This then implies the electron will be diffracted when:

$$\mathbf{k} = \frac{2\pi}{A_0}\hat{\mathbf{k}} \tag{1.58}$$

Figure 1.14 illustrates the effect that such diffraction would have on the 'free electron' curves. At wave vectors which satisfy von Laue's condition, the energy bands are disturbed

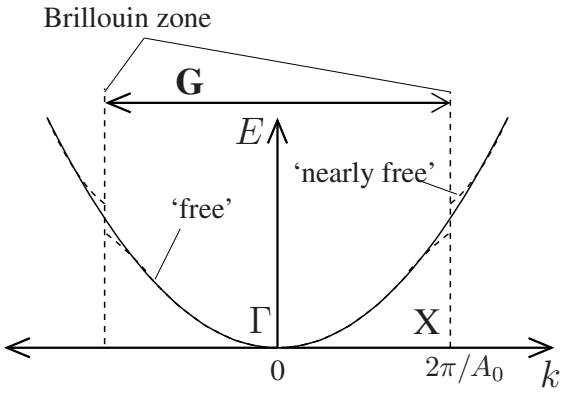


Figure 1.14: Comparison of the free and nearly free electron models

and an energy gap opens. Such an improvement on the parabolic dispersion curves of earlier is known as the *nearly free electron model*.

The space between the lowest wave vector solutions to von Laue's condition is called the first *Brillouin zone*. Note that the reciprocal lattice vectors in any particular direction span the Brillouin zone. As mentioned above, a face-centred cubic lattice has a body-centred cubic reciprocal lattice, and therefore the Brillouin zone is a three-dimensional solid, which happens to be a 'truncated octahedron' (see, for example, [5], p. 89). High-symmetry points around the Brillouin zone are often labelled for ease of reference, with the most important of these, for this work, being the $k = 0$ point, referred to as 'Γ', and the $\langle 001 \rangle$ zone edges, which are called the 'X' points.

Exercises

All necessary material parameters are given in Appendix A.

(1). Consider a free electron with a velocity of $2 \times 10^5\,\mathrm{m\,s^{-1}}$ in the z-direction.

- (a) Calculate the de Broglie wavelength and wave vector of the electron.
- (b) Assuming a wave function with the form $\psi(\mathbf{r}, t) = A \exp[i(\mathbf{k}_\bullet \mathbf{r} - \omega t)]$, where A is a constant, determine the kinetic energy of the electron using equation (1.17).
- (c) Confirm that the result is in agreement with the classical expression for kinetic energy, given in equation (1.10).

(2).
- (a) List the atomic locations that lie within a face-centred cubic unit cell of GaAs $[0 \leq (x, y, z) \leq A_0]$.
- (b) Hence, estimate the number of atoms in a cube of GaAs with $1\,\mu\mathrm{m}^3$ volume.
- (c) Taking the molar mass of a GaAs molecule as $144.645\,\mathrm{g\,mol^{-1}}$ and Avogadro's number $N_\mathrm{A} = 6.022 \times 10^{23}\,\mathrm{mol^{-1}}$, use the answer to part (b) to estimate the mass density of a GaAs crystal.

(d) Compare the calculated density with that given in Appendix A. What conclusions may be drawn from the result?

(3). Find the difference in energy between an electron and a heavy hole with a wave vector of $\mathbf{k} = 0.05\,\overrightarrow{\Gamma \mathrm{X}}$ in GaAs.

(4). Sketch the one-dimensional potential profile for the conduction- and valence-band edges in a heterostructure consisting of a $2\,\mathrm{nm}$ thick layer of GaAs surrounded by a pair of 10 nm thick layers of $\mathrm{Al}_{0.15}\mathrm{Ga}_{0.85}\mathrm{As}$.

(5). The 'L' symmetry points for a face-centred cubic crystal lie at the $\langle 111 \rangle$ zone edges. Determine their coordinates using equation (1.56) and the results in Table 1.1.

(6). Most semiconductor materials take the diamond or zinc blende structure of two interlocking face-centred cubic crystal lattices. However, GaN-based semiconductors, which are being used increasingly in green and blue laser diodes, form a hexagonal, wurtzite structure. Find the primitive lattice vectors for a hexagonal Bravais lattice and the atomic basis necessary to describe GaN.

References

[1] R. T. Weidner and R. L. Sells, *Elementary Modern Physics*, Allyn and Bacon, Boston, Third edition, 1980.

[2] P. A. M. Dirac, *The Principles of Quantum Mechanics*, Clarendon Press, Oxford, Fourth edition, 1967.

[3] S. Nakamura and G. Fasol, *The Blue Laser Diode*, Springer, Berlin, 1997.

[4] J. S. Blakemore, *Solid State Physics*, Cambridge University Press, Cambridge, Second edition, 1985.

[5] N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Saunders College Publishing, Philadelphia, 1976.

[6] M. J. Kelly, *Low Dimensional Semiconductors: Materials, Physics, Technology, Devices*, Clarendon Press, Oxford, 1995.

[7] S. Adachi, *GaAs and Related Materials: Bulk Semiconducting and Superlattice Properties*, World Scientific, Singapore, 1994.

[8] H. Landolt and R. Börnstein, Eds., *Numerical Data and Functional Relationships in Science and Technology*, vol. 22a of *Series III*, Springer-Verlag, Berlin, 1987.

[9] A. Tredicucci, C. Gmachl, F. Capasso, D. L. Sivco, and A. L. Hutchinson, 'Long wavelength superlattice quantum cascade lasers at $\lambda \approx 17\ \mu$m', *Appl. Phys. Lett.*, **74**:638, 1999.

[10] G. Bastard, 'Superlattice band structure in the envelope-function approximation', *Phys. Rev. B*, **24**(10):5693–5697, 1981.

[11] G. Bastard, *Wave Mechanics Applied to Semiconductor Heterostructures*, Monographies de physique. Halsted Press, New York, 1988.

[12] M. G. Burt, 'The justification for applying the effective-mass approximation to microstructures', *J. Phys.: Condens. Matter*, **4**:6651, 1992.

[13] M. G. Burt, 'Fundamentals of envelope function theory for electronic states and photonic modes in nanostructures', *J. Phys.: Condens. Matter*, **11**(9):53, 1999.

[14] F. Long, W. E. Hagston, and P. Harrison, 'Breakdown of the envelope function/effective mass approximation in narrow quantum wells', in *The Proceedings of the 23rd International Conference on the Physics of Semiconductors*, Singapore, 1996, pp. 1819–1822, World Scientific.

[15] Y. Hirayama, J. H. Smet, L.-H. Peng, C. G. Fonstad, and E. P. Ippen, 'Feasibility of 1.55 μm intersubband photonic devices using InGaAs/AlAs pseudomorphic quantum well structures', *Jpn. J. Appl. Phys.*, **33**(1S):890, 1994.

[16] H. Asai and Y. Kawamura, 'Intersubband absorption in $\mathrm{In}_{0.53}\mathrm{Ga}_{0.47}\mathrm{As}/\mathrm{In}_{0.52}\mathrm{Al}_{0.48}\mathrm{As}$ multiple quantum wells', *Phys. Rev. B*, **43**(6):4748–4759, 1991.

[17] A. Raymond, J. L. Robert, and C. Bernard, 'The electron effective mass in heavily doped GaAs', *J. Phys. C: Solid State Phys.*, **12**(12):2289, 1979.

[18] D. F. Nelson, R. C. Miller, and D. A. Kleinman, 'Band nonparabolicity effects in semiconductor quantum wells', *Phys. Rev. B*, **35**(14):7770–7773, 1987.

[19] M. P. Hasselbeck and P. M. Enders, 'Electron–electron interactions in the nonparabolic conduction band of narrow-gap semiconductors', *Phys. Rev. B*, **57**(16):9674–9681, 1998.

2

Solutions to Schrödinger's equation

2.1 The infinite well

The infinitely deep one-dimensional potential well is the simplest confinement potential to treat in quantum mechanics. Virtually every introductory-level text on quantum mechanics considers this system, but nonetheless it *is* worth visiting again as some of the standard assumptions, often glossed over, do have important consequences for one-dimensional confinement potentials in general.

The time-independent Schrödinger equation summarises the wave mechanics analogy to Hamilton's formulation of classical mechanics [1] for *time-independent potentials*. In essence this states that the kinetic and potential energy components sum to the total energy; in wave mechanics, these quantities are the eigenvalues of linear operators, i.e.

$$\mathcal{T}\psi + \mathcal{V}\psi = E\psi \tag{2.1}$$

where the eigenfunction ψ describes the state of the system. Again in analogy with classical mechanics, the kinetic energy operator for a particle of *constant* mass is given by:

$$\mathcal{T} = \frac{\mathcal{P}^2}{2m} \tag{2.2}$$

where $\mathcal{P}$ is the usual quantum mechanical linear momentum operator:

$$\mathcal{P} = -i\hbar\nabla = -i\hbar\left(\frac{\partial}{\partial x}\hat{\mathbf{i}} + \frac{\partial}{\partial y}\hat{\mathbf{j}} + \frac{\partial}{\partial z}\hat{\mathbf{k}}\right) \tag{2.3}$$

By using this form for the kinetic energy operator $\mathcal{T}$, the Schrödinger equation then becomes:

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\psi + V(x,y,z)\psi = E\psi \tag{2.4}$$

where the function $V(x,y,z)$ represents the potential energy of the system as a function of the spatial coordinates. Restricting this to the one-dimensional potential of interest here, the

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

Schrödinger equation for a particle of mass m in a potential well aligned along the z-axis (as in Fig. 2.1) would be:

$$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial z^2}\psi(z) + V(z)\psi(z) = E\psi(z) \tag{2.5}$$

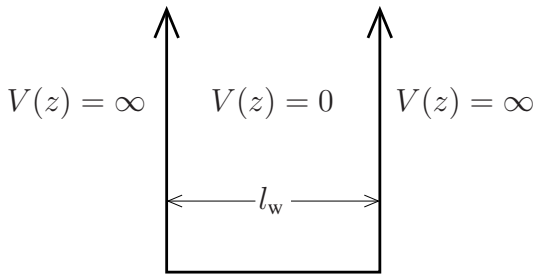


Figure 2.1: The one-dimensional infinite well confining potential

Outside of the well, $V(z) = \infty$, and hence the only possible solution is $\psi(z) = 0$, which in turn implies that all values of the energy E are allowed. Within the potential well, the Schrödinger equation simplifies to:

$$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial z^2}\psi(z) = E\psi(z) \tag{2.6}$$

which implies that the solution for ψ is a linear combination of the functions $f(z)$ which when differentiated twice give $-f(z)$. Hence try the solution:

$$\psi(z) = A\sin kz + B\cos kz \tag{2.7}$$

Substituting into equation (2.6) then gives:

$$\frac{\hbar^2 k^2}{2m}\left(A\sin kz + B\cos kz\right) = E\left(A\sin kz + B\cos kz\right) \tag{2.8}$$

$$\therefore \frac{\hbar^2 k^2}{2m} = E \tag{2.9}$$

Consideration of the boundary conditions will yield the, as yet unknown, constant k. With this aim, consider again the kinetic energy term for this system, i.e.

$$\mathcal{T} = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial z^2}\psi(z) \tag{2.10}$$

which can be rewritten as

$$\mathcal{T} = -\frac{\hbar^2}{2m}\frac{\partial}{\partial z}\left(\frac{\partial}{\partial z}\psi(z)\right) \tag{2.11}$$

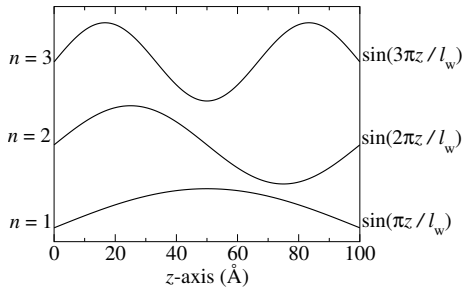


Figure 2.2: Solutions to the one-dimensional infinite well confining potential

The mathematical form of this implies that, *as a minimum*, $\psi(z)$ must be continuous. If it is not, then the first derivative will contain poles that must be avoided if the system is to have finite values for the kinetic energy. Given that $\psi(z)$ has already been deduced as zero outside of the well, then $\psi(z)$ within the well must be zero at both edges too.

If the origin is taken as the left-hand edge of the well as in Fig. 2.2, then $\psi(z)$ as defined in equation (2.8) can contain no cosine terms, i.e. $B = 0$, and hence $\psi(z) = A \sin kz$. In addition, for $\psi(0) = \psi(l_\mathrm{w}) = 0$:

$$k = \frac{\pi n}{l_\mathrm{w}} \tag{2.12}$$

where n is an integer, representing a series of solutions. Substituting into equation (2.9), the energy of the confined states is given by:

$$E_n = \frac{\hbar^2 \pi^2 n^2}{2 m l_\mathrm{w}^2} \tag{2.13}$$

The only remaining unknown is the constant factor A, which is deduced by considering the normalisation of the wave function; as $\psi^*(z)\psi(z)$ represents the probability of finding the particle at a point z, then as the particle must exist somewhere:

$$\int_0^{l_\mathrm{w}} \psi^*(z)\psi(z)\,\mathrm{d}z = 1 \tag{2.14}$$

which gives $A = \sqrt{2/l_\mathrm{w}}$, and therefore

$$\psi_n(z) = \sqrt{\frac{2}{l_\mathrm{w}}} \sin\left(\frac{\pi n z}{l_\mathrm{w}}\right) \tag{2.15}$$

Under the effective mass and envelope function approximations, the energy of an electron or hole in a hypothetical infinitely deep semiconductor quantum well can be calculated by using the effective mass m^* for the particle mass m of equation (2.13).

Figure 2.3 displays the results of calculations of the lowest three energy states of an electron in a GaAs well of width l_w surrounded by hypothetical infinite barriers (for these and all material parameters, see Appendix A). All three states show the same monotonic behaviour, with the energy decreasing as the well width increases.

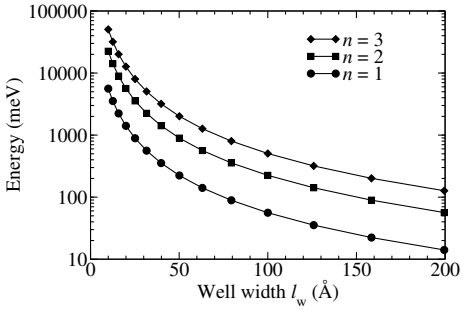


Figure 2.3: First three energy levels versus well width for an electron in a GaAs infinite potential well

The sine function solutions derived for this system are completely standard and found extensively in the literature. Although it should be noted that the arguments developed for setting the boundary conditions, i.e. $\psi(z)$ continuous, also implied that the first derivative should be continuous too, use is never made of this second boundary condition. The limitations of solution imposed by this are avoided by saying that not only is the potential infinite outside the well, but also the Schrödinger equation is not defined in these regions—a slight contradiction with the deduction of the first boundary condition. This point, i.e. that there is still ambiguity in the choice of boundary conditions for commonly accepted solutions, will be revisited later in this chapter.

2.2 In-plane dispersion

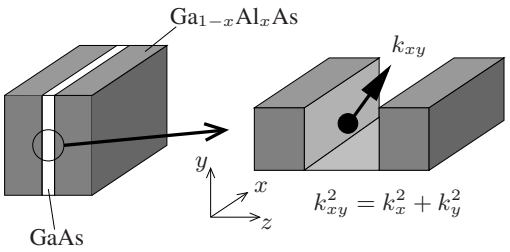


Figure 2.4: A GaAs/$Ga_{1-x}Al_xAs$ layered structure and the in-plane motion of a charge carrier

If the one-dimensional potential $V(z)$ is constructed from alternating thin layers of dissimilar semiconductors, then the particle, whether it be an electron or a hole, can move in the plane of the layers (see Fig. 2.4).

In this case, all of the terms of the kinetic energy operator are required, and hence the Schrödinger equation would be as follows:

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\psi+V(z)\psi=E\psi \tag{2.16}$$

As the potential can be written as a sum of independent functions, i.e. $V=V(x)+V(y)+V(z)$, where it just happens in this case that $V(x)=V(y)=0$, the eigenfunction of the system can be written as a product:

$$\psi(x,y,z)=\psi_x(x)\psi_y(y)\psi_z(z) \tag{2.17}$$

Using this in the above Schrödinger equation, then:

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2\psi_x}{\partial x^2}\psi_y\psi_z+\frac{\partial^2\psi_y}{\partial y^2}\psi_x\psi_z+\frac{\partial^2\psi_z}{\partial z^2}\psi_x\psi_y\right)+V(z)\psi_x\psi_y\psi_z=E\psi_x\psi_y\psi_z \tag{2.18}$$

It is then possible to identify three distinct contributions to the total energy E, one from each of the perpendicular x-, y-, and z-axes, i.e. $E=E_x+E_y+E_z$. It is said that the motions 'decouple', giving an equation of motion for each of the axes:

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi_x}{\partial x^2}\psi_y\psi_z=E_x\psi_x\psi_y\psi_z \tag{2.19}$$

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi_y}{\partial y^2}\psi_x\psi_z=E_y\psi_x\psi_y\psi_z \tag{2.20}$$

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi_z}{\partial z^2}\psi_x\psi_y+V(z)\psi_x\psi_y\psi_z=E_z\psi_x\psi_y\psi_z \tag{2.21}$$

Dividing throughout yields:

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi_x}{\partial x^2}=E_x\psi_x \tag{2.22}$$

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi_y}{\partial y^2}=E_y\psi_y \tag{2.23}$$

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi_z}{\partial z^2}+V(z)\psi_z=E_z\psi_z \tag{2.24}$$

The last component is identical to the one-dimensional Schrödinger equation for a confining potential $V(z)$ as discussed, for the particular case of an infinite well, in the last section. Consider the first and second components. Again, an eigenfunction f is sought which when differentiated twice returns $-f$; however, in this case it must be remembered that the solution will represent a moving particle. Thus the eigenfunction must reflect a current flow and have complex components, so try the standard travelling wave, $\exp(ik_x x)$. Then:

$$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\exp(ik_x x)=E_x\exp(ik_x x) \tag{2.25}$$

$$\therefore\ \frac{\hbar^2k_x^2}{2m}=E_x \tag{2.26}$$

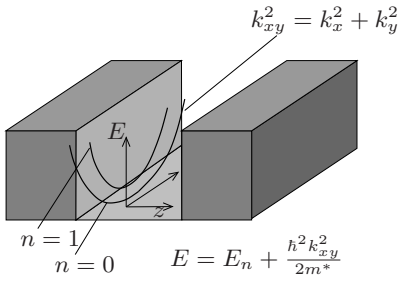


Figure 2.5: Schematic showing the in-plane ($k_{x,y}$) dispersion curves and the subband structure

which is clearly just the kinetic energy of a wave travelling along the x-axis. A similar equation follows for the y-axis, and hence the in-plane motion of a particle in a one-dimensional confining potential, but of infinite extent in the x–y plane, can be summarised as:

$$\psi_{x,y}(x,y) = \frac{1}{A}\exp\left[i(k_x x + k_y y)\right] \quad \text{and} \quad E_{x,y} = \frac{\hbar^2|\mathbf{k}_{x,y}|^2}{2m} \tag{2.27}$$

Therefore, while solutions of Schrödinger's equation along the axis of the one-dimensional potential produce discrete states of energy $E_z = E_n$, in the plane of a semiconductor quantum well there is a continuous range of allowed energies, as illustrated in Fig. 2.5. In bulk materials, such domains are called 'energy bands', while in quantum well systems these energy domains associated with confined levels are referred to as 'subbands'. Therefore the effect of the one-dimensional confining potential is to remove a degree of freedom, thus restricting the momentum of the charge carrier from three dimensions to two. It is for this reason that the states within quantum well systems are generally referred to as two-dimensional.

Later in this text, quantum wires and dots will be considered which further restrict the motion of carriers in two and three dimensions, respectively, thus giving rise to the terms one- and zero-dimensional states.

Summarising, then, within a semiconductor quantum well system the total energy of an electron or hole, of mass m^*, with in-plane momentum $k_{x,y}$, is equal to $E_z + E_{x,y}$, which is given by:

$$E = E_n + \frac{\hbar^2|\mathbf{k}_{x,y}|^2}{2m^*} \tag{2.28}$$

2.3 Extension to include band non-parabolicity

The infinite well solutions were found, so far, by approximating the dispersion of carriers as a parabola, which is reasonable for states close to the band edge. In Section 1.8, the dispersion was modelled more accurately by allowing the effective mass to be energy-dependent. The resulting effect on the energies of states in an infinite well may now be considered.

The boundary conditions for the infinite well are unaffected by band non-parabolicity, since the wave function must still reach zero as it touches the infinitely high barriers. As

such, the wave functions and wave vectors found in Section 2.1 remain valid and the result $k_n = \pi n / l_\mathrm{w}$ may be reused. The energy of the nth subband minimum is then given by substituting the appropriate wave vector into the non-parabolic dispersion relation:

$$E_n = \frac{\hbar^2 k_n^2}{2m^*(E)} = \frac{\hbar^2 \pi^2 n^2}{2m^*(E) l_\mathrm{w}^2} \tag{2.29}$$

Therefore, the energy of a subband is given by a nonlinear equation. Taking the form of the effective mass from equation (1.32) gives:

$$E_n = \frac{\hbar^2 \pi^2 n^2}{2m^*(0)(1+\alpha E_n) l_\mathrm{w}^2} \tag{2.30}$$

This can be rearranged to give the quadratic equation:

$$E_n^2 + \frac{1}{\alpha} E_n - \frac{\hbar^2 \pi^2 n^2}{2m^*(0)) l_\mathrm{w}^2 \alpha} = 0 \tag{2.31}$$

The solutions may then be written using the quadratic formula as:

$$E_n = -\frac{1}{2\alpha}\left[1 \pm \frac{1}{l_\mathrm{w}} \sqrt{\frac{2\alpha\hbar^2\pi^2 n^2 + m^*(0) l_\mathrm{w}^2}{m^*(0)}}\right] \tag{2.32}$$

Since the terms inside the square root sign are always greater than l_w^2, the negative part of the $\pm$ sign is taken in order to obtain a positive energy. Finally, this can be rearranged to give:

$$E_n = \frac{1}{2\alpha}\left[\frac{1}{l_\mathrm{w}} \sqrt{\frac{2\alpha\hbar^2\pi^2 n^2 + m^*(0) l_\mathrm{w}^2}{m^*(0)}} - 1\right] \tag{2.33}$$

Now, including the in-plane dispersion gives the total energy of a state within a subband:

$$E = E_n + \frac{\hbar^2 |\mathbf{k}_{x,y}|^2}{2m^*(0)(1+\alpha E)} \tag{2.34}$$

Multiplying by $(1+\alpha E)$ and rearranging all terms to the left-hand side gives the quadratic equation:

$$\alpha E^2 + E(1-\alpha E_n) - E_n - \frac{\hbar^2 |\mathbf{k}_{x,y}|^2}{2m^*(0)} = 0 \tag{2.35}$$

The solutions are found using the quadratic formula:

$$E = \frac{1}{2\alpha}\left[\alpha E_n - 1 \pm \sqrt{(1-\alpha E_n)^2 + 4\alpha\left(E_n + \frac{\hbar^2 |\mathbf{k}_{x,y}|^2}{2m^*(0)}\right)}\right] \tag{2.36}$$

Taking the positive solution (since the states lie above the subband minimum) and rearranging the discriminant gives:

$$E = \frac{1}{2\alpha}\left[\alpha E_n - 1 + \sqrt{(1+\alpha E_n)^2 + 4\alpha\left(\frac{\hbar^2 |\mathbf{k}_{x,y}|^2}{2m^*(0)}\right)}\right] \tag{2.37}$$

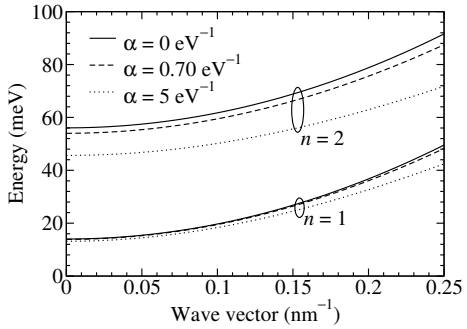


Figure 2.6: Dispersion relations for the first two subbands in a 200 Å infinite quantum well, with non-parabolicity parameters of 0, 0.70 and $5\,\mathrm{eV}^{-1}$

It is worth noting that this expression for the total energy of a state does not explicitly relate to an infinite quantum well. In fact, it is valid for any system in which carriers are confined in one dimension and free in the other two dimensions. The energy of the subband minimum E_n can be selected appropriately for the system under consideration. Figure 2.6 shows the dispersion relation (for positive wave vectors) for the first two subbands in a 200 Å infinite quantum well, with a range of non-parabolicity parameters. Two principal effects are observed as α increases. First, the energies of the subband minima (i.e. at $k = 0$) are reduced in accordance with equation (2.33). This effect is greater for higher-energy subbands. Second, the gradients of the dispersion curves are reduced. Again, this effect is greater for higher-energy subbands and also at large wave vectors.

Finally, it is worth testing the validity of the assertion in Section 1.8 that non-parabolicity effects can be neglected in many common scenarios. If attention is restricted to typical semiconductors such as GaAs ($\alpha = 0.70\,\mathrm{eV}^{-1}$), Fig. 2.6 shows that the energies of states in the first two non-parabolic subbands are close to those in the parabolic model, even at large wave vectors. In many practical semiconductor devices, only these states are of interest and hence a simple parabolic approximation is often valid. However, in some situations, for example short-wavelength (blue) lasers and LEDs, high-energy states *are* of interest and non-parabolic models may be necessary.

2.4 Density of states

It has now been shown that the original confined states within a one-dimensional potential (which could each hold two charge carriers of opposite spin) broaden into subbands, thus allowing a continuous range of carrier momenta. In order to answer the question 'Given a particular number of electrons (or holes) within a subband, what is the distribution of their energy and momenta?', the first point that is required is a knowledge of the density of states, i.e. how many electrons can exist within a range of energies. In order to answer this point for the case of subbands in quantum wells, it is necessary first to understand this property in bulk crystals.

Following the idea behind Bloch's theorem (see [2], p. 133) that an eigenstate within a

bulk semiconductor, which can be written as $\psi = (1/\Omega)\exp(i\mathbf{k}_\bullet\mathbf{r})$, must display periodicity within the lattice, then if the unit cell is of side L:

$$\psi(x,y,z) = \psi(x+L, y+L, z+L) \tag{2.38}$$

$$\therefore \psi(x,y,z) = \frac{1}{\Omega}\exp\{i[k_x(x+L)+k_y(y+L)+k_z(z+L)]\} \tag{2.39}$$

$$\therefore \psi(x,y,z) = \frac{1}{\Omega}\exp[i(k_x x + k_y y + k_z z)]\exp[i(k_x L + k_y L + k_z L)] \tag{2.40}$$

Thus for the periodicity condition to be fulfilled, the second exponential term must be identical to 1, which implies that:

$$k_x = \frac{2\pi}{L}n_x, \qquad k_y = \frac{2\pi}{L}n_y, \qquad k_z = \frac{2\pi}{L}n_z \tag{2.41}$$

where n_x, n_y and n_z are integers. Each set of values of these three integers defines a distinct state, and hence the volume of **k**-space occupied by one state is $(2\pi/L)^3$. These states fill up with successively larger values of n_x, n_y and n_z, i.e. the lowest-energy state has values (000), then permutations of (100), (110), etc., which gradually fill a sphere. At low temperatures, the sphere has a definite boundary between states that are all occupied followed by states that are unoccupied; the momentum of these states is called the *Fermi wave vector* and the equivalent energy is the *Fermi energy*. At higher temperatures, carriers near the edge of the sphere are often scattered to higher-energy states, thus 'blurring' the boundary between occupied and unoccupied states. For a more detailed description see, for example, Ashcroft and Mermin [2].

Many of the interesting phenomena associated with semiconductors derive from the properties of electrons near the Fermi energy, as it is these electrons that are able to scatter into nearby states thus changing both their energy and momenta. In order to be able to progress with descriptions of transport for example (later in this book), it is necessary to be able to describe the density of available states.

The density of states is defined as the number of states per energy per unit volume of *real* space:

$$\rho(E) = \frac{\mathrm{d}N}{\mathrm{d}E} \tag{2.42}$$

In **k**-space, the total number of states N is equal to the volume of the sphere of radius k, divided by the volume occupied by one state and divided again by the volume of real space, i.e.

$$N = 2\,\frac{4\pi k^3}{3}\,\frac{1}{(2\pi/L)^3}\,\frac{1}{L^3} \tag{2.43}$$

$$\therefore N = 2\,\frac{4\pi k^3}{3(2\pi)^3} \tag{2.44}$$

where the factor 2 has been introduced to allow for double occupancy of each state by the different carrier spins. Returning to the density of states, then:

$$\rho(E) = \frac{\mathrm{d}N}{\mathrm{d}E} = \frac{\mathrm{d}N}{\mathrm{d}k}\frac{\mathrm{d}k}{\mathrm{d}E} \tag{2.45}$$

Now equation (2.44) gives

$$\frac{\mathrm{d}N}{\mathrm{d}k} = 2\,\frac{4\pi k^2}{(2\pi)^3} = \left(\frac{k}{\pi}\right)^2 \tag{2.46}$$

In addition, the parabolic bands of effective mass theory give:

$$E = \frac{\hbar^2 k^2}{2m^*} \qquad \therefore k = \left(\frac{2m^* E}{\hbar^2}\right)^{\frac{1}{2}} \tag{2.47}$$

$$\therefore \frac{\mathrm{d}k}{\mathrm{d}E} = \left(\frac{2m^*}{\hbar^2}\right)^{\frac{1}{2}} \frac{E^{-\frac{1}{2}}}{2} \tag{2.48}$$

which finally gives the density of states in bulk as:

$$\rho(E) = \frac{1}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{\frac{3}{2}} E^{\frac{1}{2}} \tag{2.49}$$

Thus the density of states within a band, and around a minimum where the energy can be represented as a parabolic function of momentum, is a continuous function and is proportional to the square root of the energy.

2.4.1 Density-of-states effective mass

In the more general case, considered in Section 1.8, the effective mass in a semiconductor varies with respect to energy and the density of states no longer varies according to equation (2.49). Instead, the wave vector takes the form:

$$k = \left[\frac{2m^*(E)E}{\hbar^2}\right]^{\frac{1}{2}} \tag{2.50}$$

and its derivative follows from:

$$k^2 = \frac{2m^*(E)E}{\hbar^2} \quad \Rightarrow \quad \frac{\mathrm{d}k}{\mathrm{d}E} = \frac{1}{\hbar^2 k}\left[E\frac{\mathrm{d}m^*}{\mathrm{d}E} + m^*(E)\right] \tag{2.51}$$

Using equation (2.45) and $\mathrm{d}N/\mathrm{d}k = k^2/\pi^2$ from equation (2.46), the density of states:

$$\rho(E) = \frac{\mathrm{d}N}{\mathrm{d}k}\frac{\mathrm{d}k}{\mathrm{d}E} = \frac{k^2}{\pi^2}\frac{1}{\hbar^2 k}\left[E\frac{\mathrm{d}m^*}{\mathrm{d}E} + m^*(E)\right] \tag{2.52}$$

and with $k = \sqrt{2m^*(E)E}/\hbar$ then:

$$\rho(E) = \frac{\sqrt{2m^*(E)}}{\hbar^3\pi^2}E^{\frac{1}{2}}\left[E\frac{\mathrm{d}m^*}{\mathrm{d}E} + m^*(E)\right] \tag{2.53}$$

This may be rewritten in a more convenient form by making the substitution:

$$m^*_{\mathrm{d}}(E) = \left\{m^*(E)\left[E\frac{\mathrm{d}m^*}{\mathrm{d}E} + m^*(E)\right]^2\right\}^{\frac{1}{3}} \tag{2.54}$$

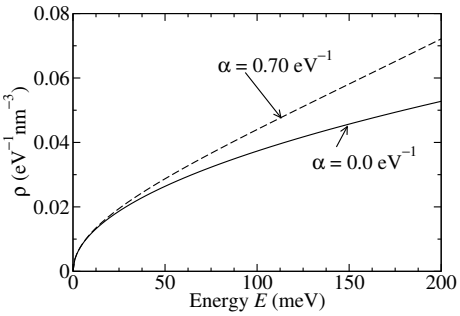


Figure 2.7: Density of states in the conduction band of bulk GaAs. The solid line shows the result when a parabolic dispersion is assumed. The dashed line corresponds to a non-parabolic dispersion with $\alpha = 1/E_{\mathrm{g}} = 0.70\,\mathrm{eV}^{-1}$

This finally gives an expression that is very similar to equation (2.49):

$$\rho(E) = \frac{1}{2\pi^2}\left(\frac{2m_{\mathrm{d}}^*(E)}{\hbar^2}\right)^{\frac{3}{2}} E^{\frac{1}{2}} \tag{2.55}$$

where the constant effective mass has been replaced by an energy-dependent property, $m_{\mathrm{d}}^*(E)$, which is known as the *density-of-states effective mass* [3].

The energy-dependent effective mass is expressed in equation (1.32) as $m^*(E) = m^*(0)(1+\alpha E)$, where $m^*(0)$ is the effective mass at the bottom of the band, α is a non-parabolicity parameter that describes the shape of the dispersion, and the energy is expressed relative to the band edge. Substituting this into equation (2.54) leads to:

$$m_{\mathrm{d}}^*(E) = m^*(0)(1+\alpha E)^{\frac{1}{3}}(1+2\alpha E)^{\frac{2}{3}} \tag{2.56}$$

in agreement with the expression given by Chakravarti and Nag [4]. Figure 2.7 shows the bulk density of states for the conduction band in GaAs both with and without the inclusion of the energy-dependent effective mass in the model. It is clear that the energy-dependent mass has little effect close to the band edge, but leads to a significant increase in the density of states at higher energies. In materials with narrow band gaps, such as InSb or InAs, the effect of an energy-dependent effective mass can be extremely significant [5], leading to densities of states many times larger than those given by equation (2.49).

2.4.2 Two-dimensional systems

The density of states in quantum well systems follows analogously; however, this time, as there are only two degrees of freedom, successive states represented by values of n_x and n_y fill a circle in $\mathbf{k}$-space, as illustrated in Fig. 2.8. Such a situation has become known as a two-dimensional electron (or hole) gas. Hence the total number of states per unit cross-sectional area is given by the spin degeneracy factor, multiplied by the area of the circle of radius k,

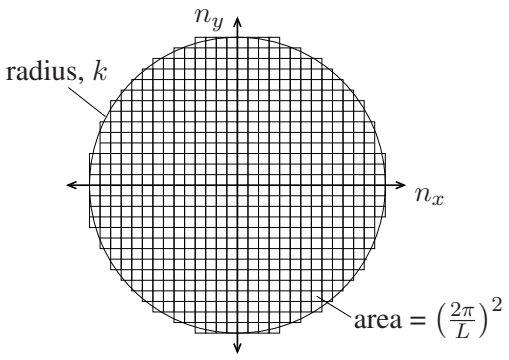


Figure 2.8: Illustration of filling the two-dimensional momenta states in a quantum well

divided by the area occupied by each state, i.e.

$$N^{2\mathrm{D}} = 2\pi k^2 \, \frac{1}{(2\pi/L)^2} \, \frac{1}{L^2} \tag{2.57}$$

$$\therefore N^{2\mathrm{D}} = 2 \, \frac{\pi k^2}{(2\pi)^2} \tag{2.58}$$

$$\therefore \frac{\mathrm{d}N^{2\mathrm{D}}}{\mathrm{d}k} = \frac{k}{\pi} \tag{2.59}$$

In analogy to the bulk three-dimensional (3D) case, define:

$$\rho^{2\mathrm{D}}(E) = \frac{\mathrm{d}N^{2\mathrm{D}}}{\mathrm{d}E} = \frac{\mathrm{d}N^{2\mathrm{D}}}{\mathrm{d}k} \, \frac{\mathrm{d}k}{\mathrm{d}E} \tag{2.60}$$

If the in-plane dispersion curves are approximated again by parabolas, then equation (2.48) can be reused as follows:

$$\rho^{2\mathrm{D}}(E) = \frac{k}{\pi} \, \left(\frac{2m^*}{\hbar^2}\right)^{\frac{1}{2}} \, \frac{E^{-\frac{1}{2}}}{2} \tag{2.61}$$

By substituting for k in terms of the energy E, using equation (2.47) then finally the density of states for a single subband in a quantum well system is given by:

$$\rho^{2\mathrm{D}}(E) = \frac{m^*}{\pi\hbar^2} \tag{2.62}$$

in agreement with Bastard [6], p. 12. This expression shows that the density of states within a subband in a quantum well is constant with respect to energy, *provided that the subband has parabolic dispersion*.

If an energy-dependent effective mass is included, as in the bulk semiconductor model, equation (2.62) is modified such that

$$\rho^{2\mathrm{D}}(E) = \frac{m^*_{\mathrm{d}}(E)}{\pi\hbar^2} \tag{2.63}$$

where the density-of-states effective mass is given for a two-dimensional system by:

$$m_{\mathrm{d}}^{*}(E) = E\frac{\mathrm{d}m^{*}}{\mathrm{d}E} + m^{*}(E) \tag{2.64}$$

Again, the energy in this expression is taken relative to the band edge. Substituting in the energy-dependent mass expression from equation (1.32) yields:

$$m_{\mathrm{d}}^{*}(E) = m^{*}(0)[1 + 2\alpha E] \tag{2.65}$$

in agreement with the result given by López–Villanueva *et al*. [3]. The energy-dependence of the effective mass therefore leads to a linear increase in the density of states with respect to energy. This effect is more significant in systems with large α and $m^{*}(0)$.

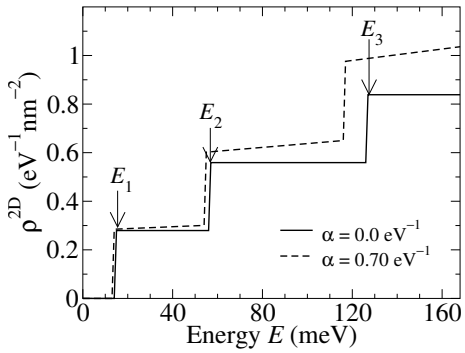


Figure 2.9: The density of states as a function of energy for a 200 Å GaAs quantum well surrounded by infinite barriers. Results are shown for both constant effective mass and energy-dependent mass, with $\alpha = 1/E_{\mathrm{g}} = 0.70\,\mathrm{eV}^{-1}$. The arrows indicate the energies of the subband minima in an infinite quantum well

If there are many (n) confined states within the quantum well system then the density of states $\rho_{\mathrm{total}}^{2\mathrm{D}}$ at any particular energy is the sum over all subbands *below* that point, which can be written succinctly as:

$$\rho_{\mathrm{total}}^{2\mathrm{D}}(E) = \rho^{2\mathrm{D}}(E)\sum_{i=1}^{n}\Theta(E - E_i) \tag{2.66}$$

where Θ is the unit step function. Figure 2.9 gives an example of the two-dimensional density of states for a particular quantum well showing the first three confined levels. Note that the steps are of equal height and occur at the subband minima—which are not equally spaced. If the effective mass is assumed to be constant with respect to energy, the density of states is constant between each pair of subbands. When band non-parabolicity is taken into account, however, the subband energies shift (as described in Section 2.3) and the density of states increases linearly between each step. The rate of increase grows with each step, since the

energy-dependent term in the density of states mass is multiplied by the number of available subbands at a given energy. As was the case in bulk semiconductors, the energy-dependence of the effective mass has a very significant effect on the density of states at higher energies, and especially in narrow-bandgap materials.

2.5 Subband populations

The total number of carriers within a subband is given by the integral of the product of the probability of occupation of a state and the density of states. Given that the carriers are fermions, then clearly the probability of occupation of a state is given by Fermi–Dirac statistics; hence:

$$N = \int_{\text{subband}} f^{\text{FD}}(E)\ \rho(E)\ \mathrm{d}E \tag{2.67}$$

where the integral is over *all* of the energies of a given subband and, of course:

$$f^{\text{FD}}(E) = \frac{1}{\exp\left[(E - E_{\text{F}})/kT\right] + 1} \tag{2.68}$$

Note that E_{F} is not the Fermi energy in the traditional sense [2]; it is a 'quasi' Fermi energy which describes the carrier population *within* a subband. For systems left to reach equilibrium, the temperature T can be assumed to be the lattice temperature; however, this is not always the case. In many quantum well devices which are subject to excitation by electrical or optical means, the 'electron temperature' can be quite different from the lattice temperature, and furthermore the subband population could be non-equilibrium and not able to be described by Fermi–Dirac statistics. For now, however, it is sufficient to discuss equilibrium electron populations and assume that the above equations are an adequate description.

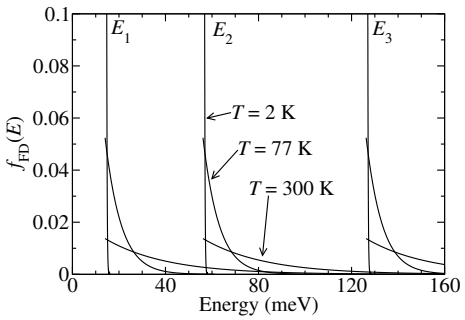


Figure 2.10: Effect of temperature on the distribution functions of the subband populations (all equal to $1 \times 10^{10}\,\text{cm}^{-2}$) of the infinite quantum well of Fig. 2.9

Given a particular number of carriers within a quantum well, which can usually be deduced directly from the surrounding doping density, it is often desirable to be able to describe

that distribution in terms of the quasi-Fermi energy E_{F}. With this aim, substitute the two-dimensional density of states appropriate to a single subband from equation (2.62) into equation (2.67). Then the carrier density, i.e. the number per unit area, is given by:

$$N = \int_{E_{\mathrm{min}}}^{E_{\mathrm{max}}} \frac{1}{\exp\left[(E - E_{\mathrm{F}})/kT\right] + 1} \; \frac{m^*}{\pi\hbar^2} \, \mathrm{d}E \tag{2.69}$$

By putting:

$$E' = \frac{E - E_{\mathrm{F}}}{kT}, \qquad \text{and then} \qquad \mathrm{d}E' = \frac{\mathrm{d}E}{kT} \tag{2.70}$$

equation (2.69) becomes:

$$N = \int_{(E_{\mathrm{min}} - E_{\mathrm{F}})/kT}^{(E_{\mathrm{max}} - E_{\mathrm{F}})/kT} \frac{kT}{\exp E' + 1} \; \frac{m^*}{\pi\hbar^2} \, \mathrm{d}E' \tag{2.71}$$

If the effective mass is taken to be a constant, which is the case for a parabolic subband, then the integral can be solved using a standard form (see, for example, Gradshteyn and Ryzhik [7], equation 2.313.2):

$$\int \frac{1}{1 + \mathrm{e}^x} \, \mathrm{d}x = x - \ln\left(1 + \mathrm{e}^x\right) \tag{2.72}$$

This may be simplified further by noting that $x = \ln(\mathrm{e}^x)$, giving:

$$\begin{aligned} \int \frac{1}{1 + \mathrm{e}^x} \, \mathrm{d}x &= \ln(\mathrm{e}^x) - \ln\left(1 + \mathrm{e}^x\right) \\ &= \ln\left(\frac{1}{1 + \mathrm{e}^{-x}}\right) \end{aligned} \tag{2.73}$$

Hence:

$$N = \frac{m^* kT}{\pi\hbar^2} \left[\ln\left(\frac{1}{1 + \mathrm{e}^{-E'}}\right) \right]_{(E_{\mathrm{min}} - E_{\mathrm{F}})/kT}^{(E_{\mathrm{max}} - E_{\mathrm{F}})/kT} \tag{2.74}$$

Evaluation then gives:

$$N = \frac{m^* kT}{\pi\hbar^2} \ln\left\{ \frac{1 + \exp\left[-(E_{\mathrm{min}} - E_{\mathrm{F}})/kT\right]}{1 + \exp\left[-(E_{\mathrm{max}} - E_{\mathrm{F}})/kT\right]} \right\} \tag{2.75}$$

The minimum of integration E_{min} is taken as the subband minima and the maximum E_{max} can be taken as a fixed cut-off energy (such as the top of the well) beyond which all carriers are assumed to be transported into other subbands. Alternatively, if E_{max} is much greater than E_{F}, the exponential term on the denominator of equation (2.75) vanishes, leaving:

$$N = \frac{m^* kT}{\pi\hbar^2} \ln\left\{1 + \exp\left[-(E_{\mathrm{min}} - E_{\mathrm{F}})/kT\right]\right\} \tag{2.76}$$

This form is simpler mathematically and is accurate at low temperatures, when the carrier density is low, or when the subbands are widely separated from each other in energy or space.

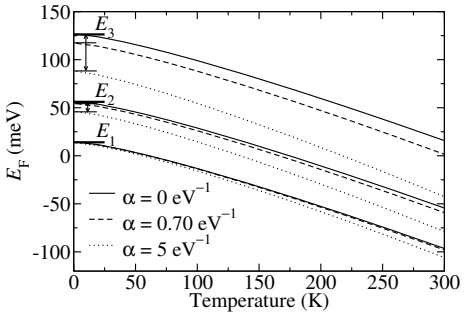


Figure 2.11: Effect of temperature on the quasi-Fermi energy describing the electron distribution of the first three states in an infinite quantum well. Results are shown for parabolic bands (solid lines), and non-parabolic bands using $\alpha = 1/E_{\mathrm{g}}$=0.70 eV^{-1} (dashed lines) and 5 eV^{-1} (dotted lines)

2.5.1 Populations in non-parabolic subbands

Strictly, equation (2.69) should make use of the density-of-states effective mass (as described in Section 2.4) rather than assuming a constant value m^*. Making use of the effective mass from equation (2.65) gives:

$$N = \frac{m^*(0)}{\pi\hbar^2} \int_{E_{\mathrm{min}}}^{E_{\mathrm{max}}} \frac{1 + 2\alpha E}{\exp\left[(E - E_{\mathrm{F}})/kT\right] + 1} \ \mathrm{d}E \tag{2.77}$$

This is more complicated to solve than equation (2.69), but an analytical solution can be found by substituting E' as above:

$$N = \frac{m^*(0)}{\pi\hbar^2} \int_{(E_{\mathrm{min}} - E_{\mathrm{F}})/kT}^{(E_{\mathrm{max}} - E_{\mathrm{F}})/kT} \frac{1 + 2\alpha(E'kT + E_{\mathrm{F}})}{\exp\left(E'\right) + 1} \ \mathrm{d}E' \tag{2.78}$$

If the upper limit is taken as infinity and the lower limit is written as $\beta = (E_{\mathrm{min}} - E_{\mathrm{F}})/kT$, this integral has the closed-form solution:

$$N = \frac{m^*(0)kT}{\pi\hbar^2} \left\{(1 + 2\alpha E_{\mathrm{F}}) \ln(1 + \mathrm{e}^{-\beta}) + F(E_{\mathrm{F}}, T, \beta)\right\} \tag{2.79}$$

where:

$$F(E_{\mathrm{F}}, T, \beta) = 2\alpha kT \left[\frac{1}{2}(E')^2 - \mathrm{Li}_2[-\exp(E')] - E' \ln[\exp(E') + 1]\right]_{\beta}^{\infty} \tag{2.80}$$

This expression makes use of the dilogarithm function, $\mathrm{Li}_2(x)$, which is widely available in mathematical reference tables and free software libraries, such as the *GNU Scientific Library* [8], or can be computed as a series expansion:

$$\mathrm{Li}_2(x) = \int_x^0 \frac{\ln(1 - t)\ \mathrm{d}t}{t} = \sum_{k=1}^{\infty} \frac{x^k}{k^2} \tag{2.81}$$

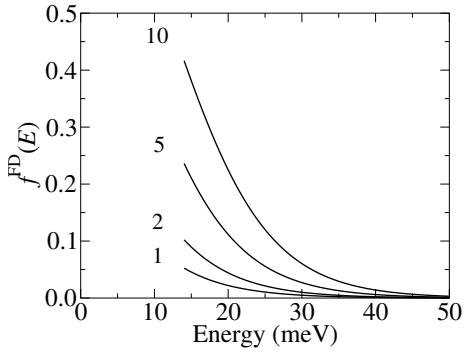


Figure 2.12: Effect of electron density, N = 1, 2, 5, 10 ($\times 10^{10}$) cm^{-2}, on the distribution function of the lowest subband of the infinite quantum well of Fig. 2.9

The upper limit in the definition of $F(E_{\mathrm{F}}, T, \beta)$ can be found by expanding the Maclaurin series of the dilogarithm term to obtain:

$$\lim_{x\to\infty} \mathrm{Li}_2(-\mathrm{e}^x) = -\frac{x^2}{2} - \frac{\pi^2}{6} \tag{2.82}$$

Noting also the limit of the final term in $F(E_{\mathrm{F}}, T, \beta)$:

$$\lim_{x\to\infty} x \ln(\mathrm{e}^x + 1) = x^2 \tag{2.83}$$

this finally yields:

$$F(E_{\mathrm{F}}, T, \beta) = 2\alpha kT \left[\frac{\pi^2}{6} + \mathrm{Li}_2(-\mathrm{e}^{\beta}) + \beta \ln\left(\mathrm{e}^{\beta} + 1\right) - \frac{\beta^2}{2}\right] \tag{2.84}$$

2.5.2 Calculation of quasi-Fermi energy

Given a total carrier density N, the quasi-Fermi energy E_{F} is the only unknown in equations (2.75) and (2.79) and can be found with standard numerical techniques. In the case where E_{max} is very large compared with the Fermi energy, and the effective mass is assumed to be constant with respect to energy, equation (2.76) can be rearranged to give the analytical solution:

$$E_{\mathrm{F}} = E_{\mathrm{min}} + kT \ln\left[\exp\left(\frac{N\pi\hbar^2}{m^* kT}\right) - 1\right] \tag{2.85}$$

Figure 2.10 gives an example of the distribution functions $f^{\mathrm{FD}}(E)$ for the first three confined levels within a 200 Å GaAs infinite quantum well when the subbands are treated as being parabolic (α = 0 eV^{-1}). As the densities of carriers, in this case electrons, have been taken as being equal and of value 1×10^{10} cm^{-2}, the distribution functions are all identical, but offset along the energy axis by the confinement energies. As mentioned above, at low temperatures the carriers tend to occupy the lowest available states, and hence the transition

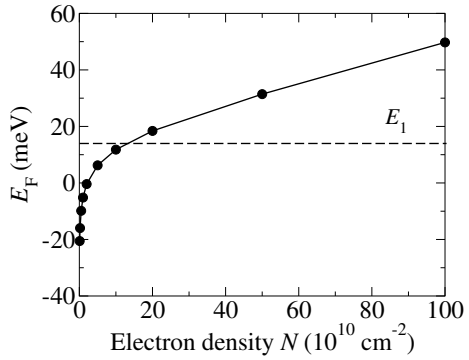


Figure 2.13: Effect of electron density on the quasi-Fermi energy describing the distribution of the ground state E_1

from states that are all occupied to those that are unoccupied is rapid—as illustrated by the 2 K data for all three subbands. As the temperature increases, the distributions broaden and a range of energies exist in which the states are partially filled, as can be seen by the 77 K and 300 K data. Physically this broadening occurs due mainly to the increase in electron–phonon scattering as the phonon population increases with temperature (more of this in Chapter 10).

Figure 2.11 displays the Fermi energy E_F as a function of temperature T for the first three subbands in the well. At low temperatures, E_F is just above the confinement energy of each subband, since the electron density is fairly low (1×10^{10} cm^{-2}). As the temperature increases, E_F falls quite markedly and below the subband minima. If this seems counterintuitive, it must be remembered that E_F is a *quasi*-Fermi energy whose only physical meaning is to describe the population within a subband—it is not the true Fermi energy of the complete system. When non-parabolicity effects are included, the quasi-Fermi level for each subband decreases slightly more rapidly with respect to temperature. It can be seen that this effect is fairly weak, and in GaAs (α = 0.70 eV^{-1}) is almost negligible. This is because the increase in density of states applies only to the weakly populated states at higher energies. However, when larger non-parabolicity parameters are used, as is the case in narrow-bandgap materials, the effect can be significant at high temperatures or in high-energy subbands.

Figure 2.12 displays the distribution functions for a range of carrier densities, for this same ground state and at a lattice temperature of 77 K. Although not obvious from the mathematics, $f^{FD}(E)$ at any particular energy E appears to scale with N. The corresponding Fermi energy is illustrated in Fig. 2.13. Clearly, the Fermi energy starts below the subband minima at this mid-range temperature, as discussed above, and, as expected, increasing numbers of carriers in the subband increases the Fermi energy, i.e. the energy of the state whose probability of occupation is $1/2$.

2.6 Thermalised distributions

So far, the distribution of carriers in each subband has been assumed to be independent. In other words, the populations of subbands could be calculated one by one, using *a priori*

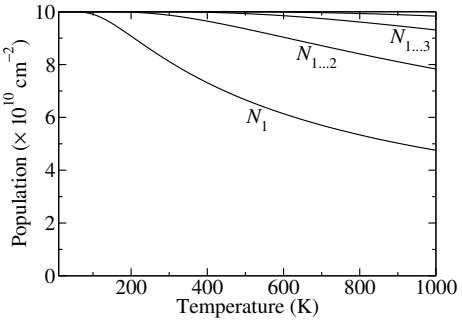


Figure 2.14: Populations of subbands in a 200 Å GaAs infinite quantum well with a total carrier density of $10 \times 10^{10}\,\mathrm{cm}^{-2}$. The lines show the *cumulative* populations up to and including the specified subband

knowledge of the quasi-Fermi energy for each subband and the carrier temperature. In reality, carriers can move between subbands, and as such, the populations of all subbands must be calculated simultaneously using techniques that take these *intersubband* transitions into account.

Chapters 10 and 12 give an overview of this type of calculation. However, it is possible to consider a much simpler model at this stage, in which the entire system is in thermal equilibrium and can be described by a single Fermi energy. This is a reasonable approximation for systems that are unaffected by external energy sources (such as light sources or electrical inputs). The total population of the system is then given by summing over all subbands:

$$N_{\text{total}} = \sum_{i=1}^{n} N(E_i, E_{\text{F}}, T) \tag{2.86}$$

where E_i is the energy of the ith subband, and the entire system is described using a single temperature T and Fermi energy E_{F}. The population N of each subband is found, as before, using either equation (2.76) or equation (2.79), depending on whether energy-dependent mass is considered.

Usually, the total population of a system is easier to specify than the Fermi energy, since this can be controlled through doping of the semiconductor material. As such, equation (2.86) can be used to find the Fermi energy that corresponds to a known population by iterative methods. This global Fermi energy may then be used to determine the population in each of the subbands. Figure 2.14 shows the distribution of electrons between the first four subbands in a 200 Å GaAs infinite quantum well with a total carrier density of $10 \times 10^{10}\,\mathrm{cm}^{-2}$, over a wide range of temperatures. At temperatures below 100 K, only the lowest subband contains a significant number of electrons. As the temperature of the system rises, an increasing proportion of electrons is transferred into higher-energy subbands. This illustrates one of the reasons why the behaviour of semiconductor heterostructure devices changes dramatically with temperature.

It is also important to note that the proportion of carriers is always greater in lower-energy subbands, regardless of the temperature. Chapter 12 discusses how an external energy

source can be used to *pump* carriers from low-energy to high-energy subbands and create a *population inversion*. However, in the absence of external energy sources, systems will usually decay rapidly into a thermalised distribution, with carriers settling preferentially into low-energy states.

2.7 Finite well with constant mass

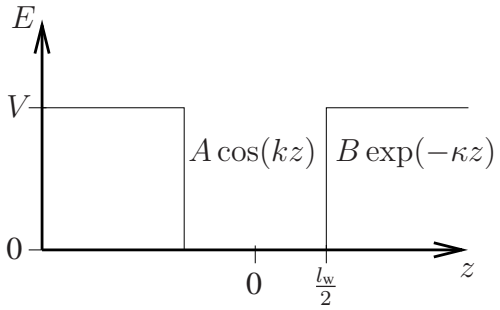


Figure 2.15: Solutions to the finite well potential

While the infinitely deep confining potential has served well as a platform for developing the physics of two-dimensional systems, more relevant to alternating layers of dissimilar semiconductors is the finite quantum well model, which under both the effective mass and envelope function approximations looks like Fig. 2.15. In particular, a layer of GaAs 'sandwiched' between two thick layers of $Ga_{1-x}Al_xAs$ would form a Type-I finite quantum well, where the conduction band has the appearance of Fig. 2.15, with the potential energy V representing the discontinuity in the conduction-band edge between the materials.

Again taking the simplest starting case of a constant electron mass m^* throughout the dissimilar layers, and neglecting movement within the plane of the layers, then the standard Schrödinger equation can be written for each of the semiconductor layers as follows:

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial z^2}\psi(z) + V\psi(z) = E\psi(z), \qquad z \leq -\frac{l_\mathrm{w}}{2} \tag{2.87}$$

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial z^2}\psi(z) = E\psi(z), \qquad -\frac{l_\mathrm{w}}{2} \leq z \leq \frac{l_\mathrm{w}}{2} \tag{2.88}$$

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial z^2}\psi(z) + V\psi(z) = E\psi(z), \qquad +\frac{l_\mathrm{w}}{2} \leq z \tag{2.89}$$

Considering solutions to the Schrödinger equation for the central well region, then as in the infinite well case, the general solution will be a sum of sine and cosine terms. As the potential is symmetric, the eigenstates will also have a definite symmetry, i.e. they will be either symmetric or antisymmetric. With the origin placed at the centre of the well, the

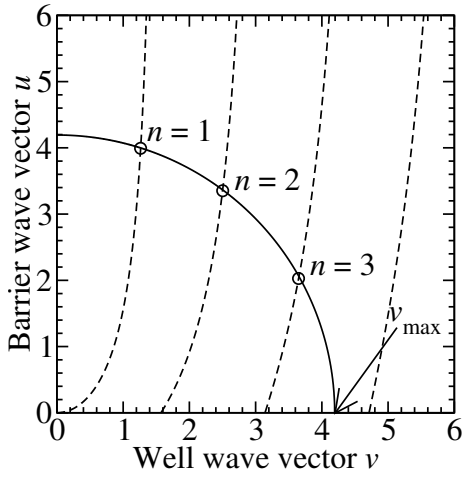


Figure 2.16: Illustration of the matching conditions for a bound state in a finite GaAs square well of width 200 Å surrounded by $100\,\mathrm{meV}$ barriers. The left-hand side of equation (2.101) is plotted as a solid line and the right-hand side is plotted as a set of dashed lines. The intersections (drawn as open circles) show the wave vectors of the bound states, and the maximum possible wave vector in the well is denoted as $v_{\max}$

symmetric (even-parity) eigenstates will then be in cosine terms, while the antisymmetric (odd-parity) states will be as sine waves.

For states confined to the well, the energy E must be less than the barrier height V, thus rearranging the Schrödinger equation for the right-hand barrier:

$$\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial z^2}\psi(z) = (V - E)\psi(z) \tag{2.90}$$

Therefore, a function f is sought which when differentiated twice gives $+f$. The exponential function fits this description, therefore consider a sum of growing $\exp(+\kappa z)$ and decaying $\exp(-\kappa z)$ exponentials. In the right-hand barrier, z is positive, and hence as z increases the growing exponential will increase too and without limit. The probability interpretation of the wave function requires that:

$$\int_{\text{all space}} \psi^*(z)\psi(z)\,\mathrm{d}z = 1 \tag{2.91}$$

which further demands that:

$$\psi(z) \to 0 \text{ and } \frac{\partial}{\partial z}\psi(z) \to 0, \qquad \text{as } z \to \pm\infty \tag{2.92}$$

These boundary conditions for states confined in wells will be used again and again and will be referred to as the *standard boundary conditions*. Using this result, the growing exponential components must be rejected and the solutions are for the even-parity states, which would

follow as:

$$\psi(z) = \begin{cases} B\exp(\kappa z), & z \leq -\frac{l_\mathrm{w}}{2} \quad (2.93a) \\ A\cos(kz), & -\frac{l_\mathrm{w}}{2} \leq z \leq \frac{l_\mathrm{w}}{2} \quad (2.93b) \\ B\exp(-\kappa z), & \frac{l_\mathrm{w}}{2} \leq z \quad (2.93c) \end{cases}$$

Note for later that these wave functions are real, and that the eigenfunctions of this confined system carry no current and hence are referred to as stationary states. Using these trial forms of the wave function in their corresponding Schrödinger equations gives the as yet unknown constants:

$$k = \frac{\sqrt{2m^* E}}{\hbar}, \qquad \kappa = \frac{\sqrt{2m^*(V-E)}}{\hbar} \tag{2.94}$$

In order to proceed, it is necessary to impose boundary conditions. Recalling the constant-mass kinetic energy operator employed in equations (2.87)–(2.89), then in order to avoid infinite kinetic energies:

$$\text{both } \psi(z) \text{ and } \frac{\partial}{\partial z}\psi(z) \text{ must be continuous}$$

Consider the interface at $z = +l_\mathrm{w}/2$. By equating ψ in the well and the barrier:

$$A\cos\left(\frac{kl_\mathrm{w}}{2}\right) = B\exp\left(-\frac{\kappa l_\mathrm{w}}{2}\right) \tag{2.95}$$

and equating the derivatives gives:

$$-kA\sin\left(\frac{kl_\mathrm{w}}{2}\right) = -\kappa B\exp\left(-\frac{\kappa l_\mathrm{w}}{2}\right) \tag{2.96}$$

Dividing equation (2.95) by equation (2.96) then gives:

$$-\frac{1}{k}\cot\left(\frac{kl_\mathrm{w}}{2}\right) = -\frac{1}{\kappa} \tag{2.97}$$

$$\therefore k\tan\left(\frac{kl_\mathrm{w}}{2}\right) - \kappa = 0 \tag{2.98}$$

Odd-parity states would require the choice of wave function in the well region as a sine wave, and hence equation (2.93b) would become $\psi = A\sin(kz)$; following through the same analysis as above gives the equation to be solved for the odd-parity eigenenergies as:

$$k\cot\left(\frac{kl_\mathrm{w}}{2}\right) + \kappa = 0 \tag{2.99}$$

Remembering that both k and κ are functions of the energy E, then equations (2.98) and (2.99) are also functions of E only. There are many ways of solving such single-variable equations, and for this particular case the literature often talks of 'graphical methods' [6, 9, 10]. Following one such method, the wave vectors and decay constants in

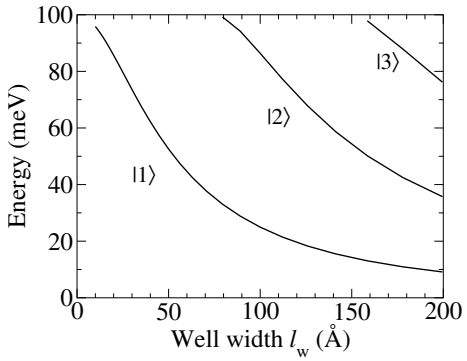


Figure 2.17: Energy levels in a GaAs single quantum well with constant effective mass $m^* = 0.067m_0$ and $V = 100\,\mathrm{meV}$

equations (2.98) and (2.99) are multiplied by $l_\mathrm{w}/2$ to obtain the dimensionless 'normalised' values, $v = kl_\mathrm{w}/2$ and $u = \kappa l_\mathrm{w}/2$, yielding a 'matching equation' for the system:

$$u = v \begin{cases} \tan v & \text{(even parity)} \\ -\cot v & \text{(odd parity)} \end{cases} \tag{2.100}$$

The expressions above can be written in terms of the variable v only, by making the further substitution $u_0^2 = v^2 + u^2$ such that:

$$\sqrt{u_0^2 - v^2} = v \begin{cases} \tan v & \text{(even parity)} \\ -\cot v & \text{(odd parity)} \end{cases} \tag{2.101}$$

Here, u_0 is a constant, whose value can be found by substituting in the definitions of u, v, k and κ:

$$\begin{aligned} u_0^2 = v^2 + u^2 &= \frac{l_\mathrm{w}^2}{4}(k^2 + \kappa^2) \\ &= \frac{2m^* l_\mathrm{w}^2}{4\hbar^2}\,(E + V - E) \\ &= \frac{m^* l_\mathrm{w}^2 V}{2\hbar^2} \end{aligned} \tag{2.102}$$

The value of u_0, therefore, contains all the structural information (i.e. effective mass, well width and barrier height) for the specific finite well system under consideration.

Equation (2.101) may now be solved by finding the values of v for which the left- and right-hand sides of the equation intersect. First, however, note that each side of the equation provides a different way of expressing the decay constant in the barrier u in terms of the wave vector in the well v. It is useful to inspect the shapes of the two functions that are described by each side of the equation.

The left-hand side of equation (2.101),

$$u = \sqrt{u_0^2 - v^2} \tag{2.103}$$

describes a circle in the u–v plane, centred around the origin of the graph, as shown by the solid line in Fig. 2.16. The radius of the circle is u_0 and, therefore, depends upon the structure of the particular system under consideration. The solutions of equation (2.101) must lie along the first quadrant of this circle, since positive wave vectors and decay constants are required. The maximum permissible values of u and v correspond to the points at which the circle meets the axes of the plot:

$$u_{\max} = v_{\max} = u_0 = \sqrt{\frac{m^* l_{\mathrm{w}}^2 V}{2\hbar^2}} \tag{2.104}$$

The right-hand side (equation (2.100)) is plotted as a branched set of dashed lines in Fig. 2.16, given by either $u = v\tan v$ or $u = -v\cot v$. Note also that that right-hand side contains no information about the specific system under consideration. As such, it represents a *general* relationship between the barrier decay constant and the well wave vector for *any* finite quantum well. The challenge, therefore, is to find the conditions under which this general relationship applies to the structure-specific information in the left-hand side of the equation.

Unfortunately, there is no analytical method for finding these intersections, and numerical methods must be used to search for the solutions instead. However, the search may be narrowed down to a finite number of well-defined ranges of v values. With this in mind, note first that the circular function can intersect each branch only once. As such, the nth branch contains all the possible solutions for the nth state in the finite quantum well. The periodicity of the $\tan$ and $\cot$ functions restricts the v value for each branch to lie in the range:

$$\frac{(n-1)\pi}{2} \leq v_n < \frac{n\pi}{2} \tag{2.105}$$

The total number of bound states, N, may be found by dividing the maximum possible normalised wave vector (equation (2.104)) by the width of a branch ($\pi/2$) and rounding this up to the nearest integer:

$$N = \left\lceil \frac{2v_{\max}}{\pi} \right\rceil . \tag{2.106}$$

A remarkable consequence of this is that *there is always at least one bound state in a finite quantum well*, regardless of the potential or the well width. This can be confirmed graphically by noting that the circle in Fig. 2.16 always intersects the first branch, even when its radius is extremely small, since the first branch passes through the origin.

For the $200\,$Å GaAs well with $100\,\mathrm{meV}$ barriers in Fig. 2.16, $v_{\max} = 4.19 = \frac{\pi}{2} \times 2.67$. Rounding this up to the nearest integer multiple of $\pi/2$ shows that there are three bound states in the system, as indicated by the small circles in the figure.

Now, each of the three solutions may be found by inspecting the respective branches in the figure over the ranges of wave vectors given by equation (2.105) for the unique value of v that satisfies equation (2.101). Widely known search techniques such as the linear bisection

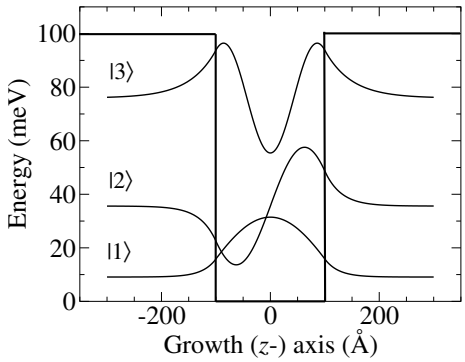


Figure 2.18: Eigenfunctions $\psi(z)$ for the first three energy levels of the 200 Å GaAs well of Fig. 2.17

method allow the appropriate value of v on each branch to be found rapidly, as discussed in Section 3.1.

Having found the normalised u and v values for each state, the absolute values are given by, $k = 2v/l_{\mathrm{w}}$ and $\kappa = 2u/l_{\mathrm{w}}$. The wave functions are then given by equation (2.93a) and the energies by equation (1.24).

All that remains now is to find the amplitudes A and B of the wave function within the well and barrier regions, respectively. The normalisation integral is used to ensure that the total probability of locating an electron over all space sums to 1. The integral expression used in equation (2.14) for an infinite well must be adapted such that its limits extend over all space, since the wave functions for a finite well leak out into the barrier regions:

$$\int_{-\infty}^{\infty} \psi^*(z)\psi(z)\,\mathrm{d}z = 1 \tag{2.107}$$

However, since the probability density function is even-symmetric about the point $z = 0$, it can be integrated from 0 to ∞ and then multiplied by 2. Substituting in the functions from equations (2.93b) and (2.93c) for the even-parity states in the system:

$$2\int_{0}^{l_{\mathrm{w}}/2} A^2\cos^2(kz)\,\mathrm{d}z + 2\int_{l_{\mathrm{w}}/2}^{\infty} B^2[\exp(-\kappa z)]^2\,\mathrm{d}z = 1 \tag{2.108}$$

Solving the integrals using standard results gives:

$$\frac{A^2}{2k}[kl_{\mathrm{w}} + \sin(kl_{\mathrm{w}})] + \frac{B^2}{\kappa}\exp(-\kappa l_{\mathrm{w}}) = 1 \tag{2.109}$$

Now, imposing the continuity condition from equation (2.95) gives:

$$B = A\exp\left(\kappa\frac{l_{\mathrm{w}}}{2}\right)\cos\left(k\frac{l_{\mathrm{w}}}{2}\right) \tag{2.110}$$

Substituting this into equation (2.109) and rearranging for A gives:

$$A = \left[\frac{l_{\mathrm{w}}}{2} + \frac{1}{2k}\sin(kl_{\mathrm{w}}) + \frac{1}{\kappa}\cos^2\left(k\frac{l_{\mathrm{w}}}{2}\right)\right]^{-\frac{1}{2}} \tag{2.111}$$

and the corresponding value of B is then obtained using equation (2.110).

A very similar derivation may be used for the odd-parity states to obtain:

$$A = \left[\frac{l_\mathrm{w}}{2} - \frac{1}{2k}\sin(kl_\mathrm{w}) + \frac{1}{\kappa}\sin^2\left(k\frac{l_\mathrm{w}}{2}\right)\right]^{-\frac{1}{2}} \tag{2.112}$$

and

$$B = A\exp\left(\kappa\frac{l_\mathrm{w}}{2}\right)\sin\left(k\frac{l_\mathrm{w}}{2}\right) \tag{2.113}$$

Figures 2.17 and 2.18 summarise the application of the method to a GaAs single quantum well, surrounded by barrier of height $100\,\mathrm{meV}$, with the same effective mass. Clearly, as the well width increases, the energy levels all decrease, with the presence of excited states being also apparent at the larger well widths. The eigenstates are labelled according to their principle quantum number (energy order). The even and odd parities of the states within the well can be seen in Fig. 2.18.

2.7.1 Unbound states

In the previous section, it was shown that a finite quantum well contains at least one *bound state* at an energy below the confining potential, and that the wave function of the bound states was always a decaying exponential function within the barrier regions. At energies *above* the confining potential, however, $E - V$ is positive, and the wave function is therefore a sinusoidal function similar to that of a free particle within both the well and the 'barrier' layers. However, the situation is different from that of a free particle, in that the amplitude and wave vector of the solution can be different within each layer. As before, the symmetry of the system dictates that the solution must have either odd or even parity. The even solutions are then given by:

$$\psi(z) = \begin{cases} B\cos\left(k_\mathrm{b}z - \phi\right), & z \leq -\frac{l_\mathrm{w}}{2} \\ A\cos\left(kz\right), & -\frac{l_\mathrm{w}}{2} \leq z \leq \frac{l_\mathrm{w}}{2} \\ B\cos\left(k_\mathrm{b}z + \phi\right), & \frac{l_\mathrm{w}}{2} \leq z \end{cases} \tag{2.114}$$

The wave vector within the 'barrier' layers is, in this case, given by:

$$k_\mathrm{b} = \frac{\sqrt{2m^*(E-V)}}{\hbar} \tag{2.115}$$

and a phase-shift, ϕ has been introduced to ensure continuity between sinusoidal functions with differing wave vectors. Now, matching the wave functions at the interfaces gives:

$$A\cos\left(\frac{kl_\mathrm{w}}{2}\right) = B\cos\left(\frac{k_\mathrm{b}l_\mathrm{w}}{2} + \phi\right) \tag{2.116}$$

and equating the derivatives gives:

$$-kA\sin\left(\frac{kl_\mathrm{w}}{2}\right) = -k_\mathrm{b}B\sin\left(\frac{k_\mathrm{b}l_\mathrm{w}}{2} + \phi\right) \tag{2.117}$$

Dividing equation (2.117) by equation (2.116) then gives:

$$k \tan\left(\frac{k l_\mathrm{w}}{2}\right) = k_\mathrm{b} \tan\left(\frac{k_\mathrm{b} l_\mathrm{w}}{2} + \phi\right) \tag{2.118}$$

Unlike the case of states below the confining potential, there are infinitely many solutions to this equation (recalling that $E > V$ in this case). In other words, a *continuum* of states exists at energies above the barriers whereas a finite set of discrete states exists at energies below the barriers. For any given energy, the wave vectors k and k_b can, therefore, be calculated and the phase constant can be found directly by rearranging equation (2.118):

$$\phi = \arctan\left[\frac{k}{k_\mathrm{b}} \tan\left(\frac{k l_\mathrm{w}}{2}\right)\right] - \frac{k_\mathrm{b} l_\mathrm{w}}{2} \tag{2.119}$$

A very similar derivation applies to the odd-parity states, in which the wave functions are:

$$\psi(z) = \begin{cases} B \sin\left(k_\mathrm{b} z - \phi\right), & z \leq -\frac{l_\mathrm{w}}{2} \\ A \sin\left(k z\right), & -\frac{l_\mathrm{w}}{2} \leq z \leq \frac{l_\mathrm{w}}{2} \\ B \sin\left(k_\mathrm{b} z + \phi\right), & \frac{l_\mathrm{w}}{2} \leq z \end{cases} \tag{2.120}$$

and the phase constant is determined as:

$$\phi = \arctan\left[\frac{k_\mathrm{b}}{k} \tan\left(\frac{k l_\mathrm{w}}{2}\right)\right] - \frac{k_\mathrm{b} l_\mathrm{w}}{2} \tag{2.121}$$

The normalisation constants, A and B, remain undefined at present, because the wave functions extend infinitely in both directions and the normalisation integral has no unique solution. However, an approximate form can be found by assuming that the entire system is confined within a very wide infinite quantum well of width l_z. In such a system, the finite well of width l_w would only represent a tiny perturbation to the overall Hamiltonian, and it follows that the result found in Section 2.1 can be reused, giving $B = 1/\sqrt{l_z}$. The amplitude of the wave function within the well region is then given by rearranging the matching equations to obtain:

$$A = B \frac{\cos\left(k_\mathrm{b} \frac{l_\mathrm{w}}{2} + \phi\right)}{\cos\left(k \frac{l_\mathrm{w}}{2}\right)} \tag{2.122}$$

and

$$A = B \frac{\sin\left(k_\mathrm{b} \frac{l_\mathrm{w}}{2} + \phi\right)}{\sin\left(k \frac{l_\mathrm{w}}{2}\right)} \tag{2.123}$$

for the even- and odd-parity states, respectively.

A remarkable property of this normalisation is that A/B becomes infinitely large for certain values of k (i.e. when the denominator of equations (2.122) or (2.123) becomes zero). In other words, the wave function becomes entirely localised within the well region at certain energies. For the even states this occurs whenever $k = (2n-1)\pi/l_\mathrm{w}$, and for the odd states whenever $k = 2n\pi/l_\mathrm{w}$. Putting these together and substituting into the expression for carrier energy shows that these *resonant continuum states* or *virtual levels* [11] lie at:

$$E = \frac{\hbar^2 \pi^2 n^2}{2 m^* l_\mathrm{w}^2} \tag{2.124}$$

i.e. the solutions for an infinite quantum well of width l_{w}, provided of course that the energies lie *above* the barrier potential. This is an intriguing property of quantum heterostructure systems: that the effects of confinement are strongly apparent, even when a carrier has greater energy than the confining potential. It is important to note, however, that the notion of carriers being 'trapped' in these virtual levels is not entirely accurate, since they can easily 'escape' into adjacent states within the continuum. Nevertheless, these virtual levels do lead to interesting resonant effects in the optical behaviour of devices (see Section 11.6).

2.7.2 *Effective mass mismatch at heterojunctions*

Quantum wells are only fabricated by forming heterojunctions between *different* semiconductors. From an electronic viewpoint, the semiconductors are different because they have different band structures. The difference in perhaps the most fundamental property of a semiconductor, i.e. the band gap (and its alignment), is accounted for by specifying a band offset, which has been labelled V. Of course, there are many other properties which are also different, such as the dielectric constant, the lattice constant and, what is considered the next most important quantity, the effective mass. It is generally accepted that the calculation of *static* energy levels within quantum wells should account for the variation in the effective mass across the heterojunction.

This problem has been continuously addressed in the literature [12, 13, 14] since the earliest work of Conley *et al*. [15] and BenDaniel and Duke [16], who derived the boundary conditions on solutions of the *envelope functions* as:

$$\text{both } \psi(z) \text{ and } \frac{1}{m^*}\frac{\partial}{\partial z}\psi(z) \text{ continuous} \tag{2.125}$$

by considering electron transport across a heterojunction. These boundary conditions have become known as the BenDaniel–Duke boundary conditions. They are a consequence of requiring that the kinetic energy operator is Hermitian, so when the mass ceases to be a constant then:

$$\mathcal{T} = -\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial z^2} \text{ needs to be transformed into } -\frac{\hbar^2}{2}\frac{\partial}{\partial z}\left(\frac{1}{m^*(z)}\frac{\partial}{\partial z}\right) \tag{2.126}$$

with the boundary conditions above arising from the fact that when $\mathcal{T}$ acts on $\psi(z)$ it must return something finite.

Therefore, applying this extension to the finite well of the previous section would require the Schrödinger equation to be specified in each region as follows:

$$-\frac{\hbar^2}{2m_{\mathrm{b}}^*}\frac{\partial^2}{\partial z^2}\psi(z) + V\psi(z) = E\psi(z), \qquad z \leq -\frac{l_{\mathrm{w}}}{2} \tag{2.127}$$

$$-\frac{\hbar^2}{2m_{\mathrm{w}}^*}\frac{\partial^2}{\partial z^2}\psi(z) = E\psi(z), \qquad -\frac{l_{\mathrm{w}}}{2} \leq z \leq \frac{l_{\mathrm{w}}}{2} \tag{2.128}$$

$$-\frac{\hbar^2}{2m_{\mathrm{b}}^*}\frac{\partial^2}{\partial z^2}\psi(z) + V\psi(z) = E\psi(z), \qquad +\frac{l_{\mathrm{w}}}{2} \leq z \tag{2.129}$$

with the additional constraint of the matching conditions of equation (2.125).

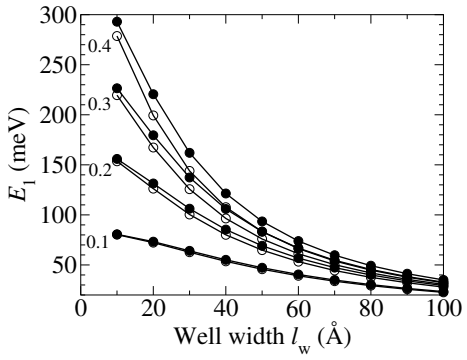


Figure 2.19: Electron ground state energy E_1 as a function of the width l_w of a GaAs well surrounded by $Ga_{1-x}Al_xAs$ barriers, calculated for both the constant-mass model (closed circles) and different barrier masses (open circles) and for a range of barrier alloy concentrations x (= 0.1, 0.2, 0.3, 0.4)

The solutions follow as previously for the constant-mass case, in equations (2.93), but now k and κ contain different effective masses:

$$k = \frac{\sqrt{2m^*_\mathrm{w}E}}{\hbar}, \qquad \kappa = \frac{\sqrt{2m^*_\mathrm{b}(V-E)}}{\hbar} \tag{2.130}$$

The method of solution is almost identical: equating the envelope functions at the interface $z = +l_\mathrm{w}/2$,

$$A\cos\left(\frac{kl_\mathrm{w}}{2}\right) = B\exp\left(-\frac{\kappa l_\mathrm{w}}{2}\right) \tag{2.131}$$

and equating $1/m^*$ times the derivative gives:

$$-\frac{kA}{m^*_\mathrm{w}}\sin\left(\frac{kl_\mathrm{w}}{2}\right) = -\frac{\kappa B}{m^*_\mathrm{b}}\exp\left(-\frac{\kappa l_\mathrm{w}}{2}\right) \tag{2.132}$$

Dividing equation (2.132) by equation (2.131), the matching equation for the even-parity states is:

$$\frac{k}{m^*_\mathrm{w}}\tan\left(\frac{kl_\mathrm{w}}{2}\right) - \frac{\kappa}{m^*_\mathrm{b}} = 0 \tag{2.133}$$

and similarly for the odd-parity solutions:

$$\frac{k}{m^*_\mathrm{w}}\cot\left(\frac{kl_\mathrm{w}}{2}\right) + \frac{\kappa}{m^*_\mathrm{b}} = 0 \tag{2.134}$$

and obviously, equating the effective masses gives the original constant-mass equations.

As before, k and κ are replaced by normalised constants u and v. However, the effective mass mismatch must be included in this substitution. For convenience, the same definition as before is used for $v = kl_\mathrm{w}/2$ and the mass ratio is included in the other substitution variable:

$$u = \frac{\kappa l_\mathrm{w}}{2}\frac{m^*_\mathrm{w}}{m^*_\mathrm{b}} \tag{2.135}$$

These substitutions yield the same matching equation (2.101) as before. In this case, however, the constant u_0 is modified to become a variable that contains the mass ratio:

$$\begin{aligned} u_0^2 = v^2 + u^2 &= \frac{l_\mathrm{w}^2}{4}\left[k^2 + \kappa^2\left(\frac{m_\mathrm{w}^*}{m_\mathrm{b}^*}\right)^2\right] \\ &= \frac{l_\mathrm{w}^2}{4\hbar^2}\left[2m_\mathrm{w}^* E + 2m_\mathrm{b}^*(V-E)\left(\frac{m_\mathrm{w}^*}{m_\mathrm{b}^*}\right)^2\right] \\ &= \frac{m_\mathrm{w}^* l_\mathrm{w}^2}{2\hbar^2}\left[\frac{m_\mathrm{w}^*}{m_\mathrm{b}^*}V + E\left(1 - \frac{m_\mathrm{w}^*}{m_\mathrm{b}^*}\right)\right] \end{aligned} \tag{2.136}$$

The left-hand side of equation (2.101) again contains all of the structural information about the specific system, while the right-hand side is exactly the same as in the constant-mass case. The dependence of u_0 on energy means that the left-hand side of equation (2.101) is no longer a perfect circle. However, the maximum permissible value of v may still be located and the same solution method may be used as before, in which each branch of the right-hand side is inspected for the respective solution of the Schrödinger equation.

To identify v_max note that this, as before, occurs when the left-hand side of equation (2.101) intersects the v-axis. At this point, u, and hence κ, must be zero. It follows from equation (2.130) that $E = V$ at this point. Substituting this into the definition of v gives:

$$v_\mathrm{max} = \frac{k l_\mathrm{w}}{2} = \frac{\sqrt{2m_\mathrm{w}^* V}\, l_\mathrm{w}}{2\hbar} = \sqrt{\frac{m_\mathrm{w}^* l_\mathrm{w}^2 V}{2\hbar^2}} \tag{2.137}$$

which is similar to equation (2.104), as expected.

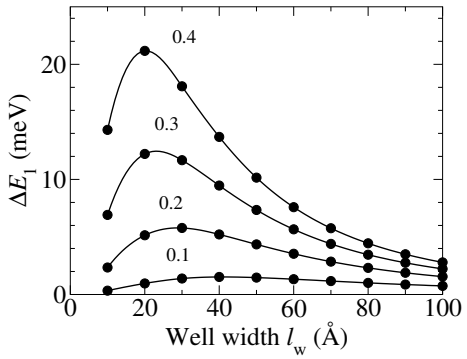


Figure 2.20: Energy difference $\Delta E_1 = E_1(m^*\text{=constant}) - E_1(m^*(z))$ for the structures shown in Fig. 2.19

Figure 2.19 compares the electron ground state energy calculated with a constant GaAs mass, with the energy calculated with a material-dependent mass, for a single GaAs quantum well surrounded by $\mathrm{Ga}_{1-x}\mathrm{Al}_x\mathrm{As}$ barriers. All of the calculated ground state energies E_1

decrease with increasing well width and increase with increasing Al fraction in the barriers (barrier height $V \propto x$). The effective mass in $Ga_{1-x}Al_xAs$ is greater than in GaAs, hence the variable mass calculations give energies less than the constant-mass model for *all* systems considered here (see later).

Figure 2.20 displays the calculated ground state energy difference between the two models, $\Delta E_1 = E_1(m^*\text{=constant}) - E_1(m^*(z))$. Clearly, and as would be expected, the larger the difference in the effective masses between the materials, the larger the difference in ground state energies.

The unbound states above the potential barriers may also be deduced using the same procedure as above [11]. The phase constants (equations (2.119) and (2.121)) are modified to be:

$$\phi = \begin{cases} \arctan\left[\frac{km_{\mathrm{b}}^*}{k_{\mathrm{b}}m_{\mathrm{w}}^*}\tan\left(\frac{kl_{\mathrm{w}}}{2}\right)\right] - \frac{k_{\mathrm{b}}l_{\mathrm{w}}}{2}, & \text{even} \\ \arctan\left[\frac{k_{\mathrm{b}}m_{\mathrm{w}}^*}{km_{\mathrm{b}}^*}\tan\left(\frac{kl_{\mathrm{w}}}{2}\right)\right] - \frac{k_{\mathrm{b}}l_{\mathrm{w}}}{2}, & \text{odd} \end{cases} \tag{2.138}$$

2.7.3 The infinite barrier height and mass limits

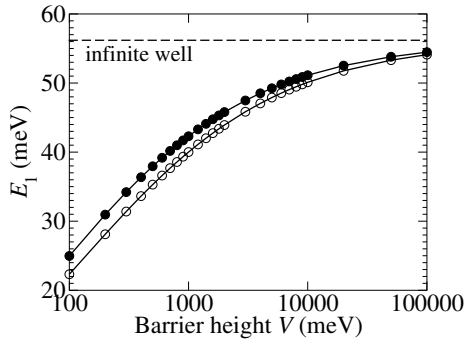


Figure 2.21: Electron ground state energy E_1 as a function of barrier height V, for a 100 Å GaAs finite well with constant mass (closed circles) and different barrier mass (fixed at mass in $Ga_{0.6}Al_{0.4}As$ (open circles))

It is interesting to take theoretical models to certain limits as a means of verifying, or otherwise, their behaviour with what might be expected intuitively. Figure 2.21 illustrates that, in the limit of large barrier heights V, the finite well model recovers the result of the infinite well model, which is what would be hoped for, thus increasing confidence in the derivation.

As has been found in the literature, Fig. 2.22 illustrates the results of allowing the barrier mass to increase without limit, while keeping all other parameters constant. The tendency for the ground state energy to tend towards zero and the unusual-looking wave functions

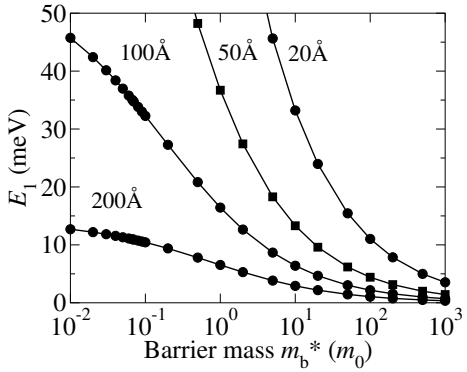


Figure 2.22: Electron ground state energy E_1 as a function of the mass in the barrier, for a variety of GaAs well widths; the barrier height is fixed at that for $Ga_{0.6}Al_{0.4}As$

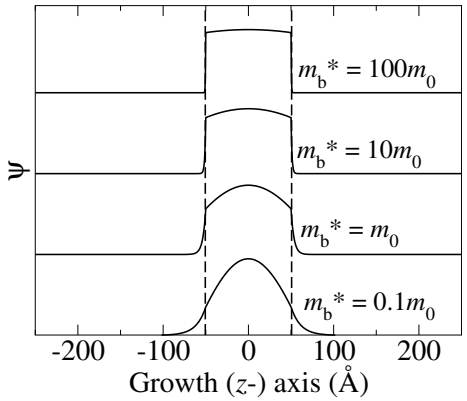


Figure 2.23: Electron ground state wave functions for several barrier masses, as given, for a 100 Å GaAs well with a barrier height fixed at that for $Ga_{0.6}Al_{0.4}As$

of Fig. 2.23 have been well documented [6] and are a direct consequence of the second boundary condition, ψ'/m^*.

It is worthwhile considering this limit still further. Figure 2.22 clearly shows that the ground state energy tends to zero for all well widths:

$$\lim_{m_b \to \infty} E_1 = 0, \qquad \forall\, l_w \tag{2.139}$$

which at first sight appears to violate Heisenberg's uncertainty principle, in that the ground state energy *can* be forced to zero for infinitesimal well widths, and thus the error in the measurement of position and momentum can be made arbitrarily small. This was an argument advanced by Hagston *et al*. [17]; however, direct evaluation of the variance in the position and momentum of these states will show that the uncertainty relationships are not violated. At the

moment, the techniques for such a calculation have not been covered; hence such discussions will be returned to later in Section 3.21.

2.8 Extension to multiple-well systems

For a whole variety of reasons, semiconductor physicists and electronic device engineers need to design and fabricate heterostructures more complex than the single quantum well [10]. The above techniques of solving Schrödinger's equation in each semiconductor layer separately and deducing the unknown coefficients by implementing boundary conditions can be applied to these multilayered systems. In particular, consider the asymmetric double quantum well of Fig. 2.24.

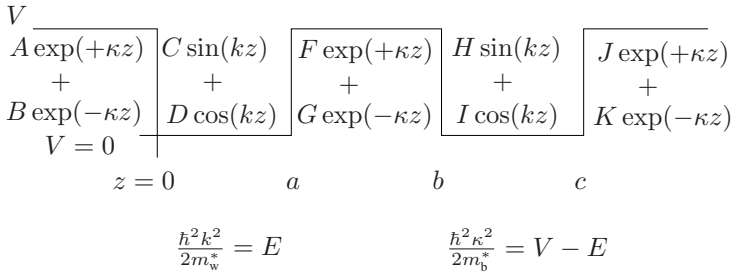


Figure 2.24: Solutions to Schrödinger's equation in a double quantum well

Choosing both wells to be of the same depth, and all of the barriers to be of the same height V, allows the simplification that k and κ are constant throughout the structure; the general solutions to the Schrödinger equation in each region are as given in the figure. Proceeding as before, then matching envelope functions and the derivative divided by the mass at the interfaces, considering $z = 0$ first:

$$A + B = D \tag{2.140}$$

$$\frac{1}{m^*_{\mathrm{b}}}(\kappa A - \kappa B) = \frac{1}{m^*_{\mathrm{w}}} kC \tag{2.141}$$

Then with $z = a$:

$$C \sin ka + D \cos ka = F \exp(+\kappa a) + G \exp(-\kappa a) \tag{2.142}$$

$$\frac{1}{m^*_{\mathrm{w}}}(kC \cos ka - kD \sin ka) = \frac{1}{m^*_{\mathrm{b}}}[\kappa F \exp(+\kappa a) - \kappa G \exp(-\kappa a)] \tag{2.143}$$

Next, with $z = b$:

$$F \exp(+\kappa b) + G \exp(-\kappa b) = H \sin kb + I \cos kb \tag{2.144}$$

$$\frac{1}{m^*_{\mathrm{b}}}[\kappa F \exp(+\kappa b) - \kappa G \exp(-\kappa b)] = \frac{1}{m^*_{\mathrm{w}}}(kH \cos kb - kI \sin kb) \tag{2.145}$$

Finally, with $z = c$:

$$H \sin kc + I \cos kc = J \exp(+\kappa c) + K \exp(-\kappa c) \tag{2.146}$$

$$\frac{1}{m_{\mathrm{w}}^*}(kH \cos kc - kI \sin kc) = \frac{1}{m_{\mathrm{b}}^*}[\kappa J \exp(+\kappa c) - \kappa K \exp(-\kappa c)] \tag{2.147}$$

All of this can be rewritten more neatly in matrix form as:

$$\begin{pmatrix} 1 & 1 \\ \frac{1}{m_{\mathrm{b}}^*}\kappa & -\frac{1}{m_{\mathrm{b}}^*}\kappa \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ \frac{1}{m_{\mathrm{w}}^*}k & 0 \end{pmatrix} \begin{pmatrix} C \\ D \end{pmatrix} \tag{2.148}$$

$$\begin{pmatrix} \sin ka & \cos ka \\ \frac{k}{m_{\mathrm{w}}^*}\cos ka & -\frac{k}{m_{\mathrm{w}}^*}\sin ka \end{pmatrix} \begin{pmatrix} C \\ D \end{pmatrix} = \begin{pmatrix} \mathrm{e}^{+\kappa a} & \mathrm{e}^{-\kappa a} \\ \frac{\kappa}{m_{\mathrm{b}}^*}\mathrm{e}^{+\kappa a} & -\frac{\kappa}{m_{\mathrm{b}}^*}\mathrm{e}^{-\kappa a} \end{pmatrix} \begin{pmatrix} F \\ G \end{pmatrix} \tag{2.149}$$

$$\begin{pmatrix} \mathrm{e}^{+\kappa b} & \mathrm{e}^{-\kappa b} \\ \frac{\kappa}{m_{\mathrm{b}}^*}\mathrm{e}^{+\kappa b} & -\frac{\kappa}{m_{\mathrm{b}}^*}\mathrm{e}^{-\kappa b} \end{pmatrix} \begin{pmatrix} F \\ G \end{pmatrix} = \begin{pmatrix} \sin kb & \cos kb \\ \frac{k}{m_{\mathrm{w}}^*}\cos kb & -\frac{k}{m_{\mathrm{w}}^*}\sin kb \end{pmatrix} \begin{pmatrix} H \\ I \end{pmatrix} \tag{2.150}$$

$$\begin{pmatrix} \sin kc & \cos kc \\ \frac{k}{m_{\mathrm{w}}^*}\cos kc & -\frac{k}{m_{\mathrm{w}}^*}\sin kc \end{pmatrix} \begin{pmatrix} H \\ I \end{pmatrix} = \begin{pmatrix} \mathrm{e}{+\kappa c} & \mathrm{e}^{-\kappa c} \\ \frac{\kappa}{m_{\mathrm{b}}^*}\mathrm{e}^{+\kappa c} & -\frac{\kappa}{m_{\mathrm{b}}^*}\mathrm{e}^{-\kappa c} \end{pmatrix} \begin{pmatrix} J \\ K \end{pmatrix} \tag{2.151}$$

Labelling the 2×2 matrix for the left-hand side of the nth interface as M_{2n-1} and the corresponding matrix for the right-hand side of the interface as M_{2n}, $n = 1, 2, 3, \ldots$, then the above matrix equations would become:

$$\mathsf{M}_1 \begin{pmatrix} A \\ B \end{pmatrix} = \mathsf{M}_2 \begin{pmatrix} C \\ D \end{pmatrix} \tag{2.152}$$

$$\mathsf{M}_3 \begin{pmatrix} C \\ D \end{pmatrix} = \mathsf{M}_4 \begin{pmatrix} F \\ G \end{pmatrix} \tag{2.153}$$

$$\mathsf{M}_5 \begin{pmatrix} F \\ G \end{pmatrix} = \mathsf{M}_6 \begin{pmatrix} H \\ I \end{pmatrix} \tag{2.154}$$

$$\mathsf{M}_7 \begin{pmatrix} H \\ I \end{pmatrix} = \mathsf{M}_8 \begin{pmatrix} J \\ K \end{pmatrix} \tag{2.155}$$

Now equation (2.152) gives:

$$\begin{pmatrix} A \\ B \end{pmatrix} = \mathsf{M}_1^{-1}\mathsf{M}_2 \begin{pmatrix} C \\ D \end{pmatrix} \tag{2.156}$$

and equation (2.153) gives:

$$\begin{pmatrix} C \\ D \end{pmatrix} = \mathsf{M}_3^{-1}\mathsf{M}_4 \begin{pmatrix} F \\ G \end{pmatrix} \tag{2.157}$$

$$\therefore \begin{pmatrix} A \\ B \end{pmatrix} = \mathsf{M}_1^{-1}\mathsf{M}_2\mathsf{M}_3^{-1}\mathsf{M}_4 \begin{pmatrix} F \\ G \end{pmatrix} \tag{2.158}$$

and eventually:

$$\begin{pmatrix} A \\ B \end{pmatrix} = \mathsf{M}_1^{-1}\mathsf{M}_2\mathsf{M}_3^{-1}\mathsf{M}_4\mathsf{M}_5^{-1}\mathsf{M}_6\mathsf{M}_7^{-1}\mathsf{M}_8 \begin{pmatrix} J \\ K \end{pmatrix} \tag{2.159}$$

The product of the eight 2×2 matrices is still a 2×2 matrix; thus writing:

$$\begin{pmatrix} A \\ B \end{pmatrix} = \mathsf{M} \begin{pmatrix} J \\ K \end{pmatrix} \tag{2.160}$$

then

$$A = M_{11}J + M_{12}K \tag{2.161}$$

$$B = M_{21}J + M_{22}K \tag{2.162}$$

Again, the probability interpretation of the wave function implies that the wave function must tend towards zero into the outer barriers, i.e. the coefficients of the growing exponentials must be zero. *In this case*, with the origin at the first interface (see Fig. 2.24), this implies that $B = 0$ and $J = 0$, and hence the second of the above equations would imply that $M_{22} = 0$. As all of the elements of M are functions of both k and κ, which are both in turn functions of the energy E, then an energy is sought which satisfies:

$$M_{22}(E) = 0 \tag{2.163}$$

which can be found by standard numerical procedures, such as Newton–Raphson iteration as discussed in Section 3.2. This approach and variations upon it are often referred to as the *transfer matrix technique*. Once the energy is known, the coefficients $A, \ldots, K$ follow simply and the envelope wave function can be deduced.

2.9 The asymmetric single quantum well

The above system has illustrated a method of solution for a system in which the confinement energy E was always less than the barrier height V. While this is often true, a class of quantum well structures exists in which there is more than one barrier height (or well depth); Fig. 2.25 illustrates one such system.

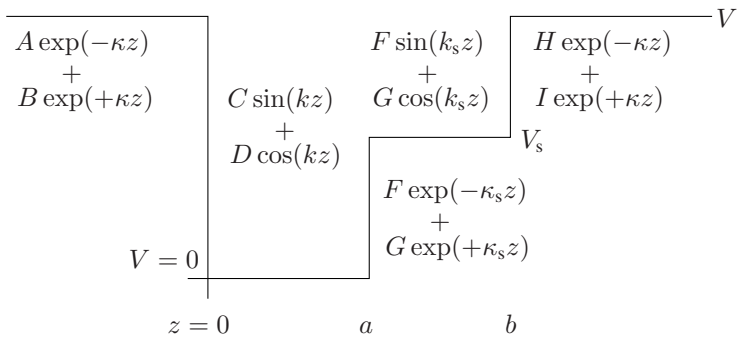


Figure 2.25: Solutions to Schrödinger's equation in a stepped asymmetric quantum well

A confined state could exist that has an energy below or above that of the step height V_s, hence the functional form of the solution to Schrödinger's equation in that region is dependent

upon the energy E. This is illustrated schematically in Fig. 2.25. For states above the step, the wave function has the form of a normal well state, but with a k value different from that of the well region. For states of energy below the step potential V_s, the wave function resembles a barrier state. This causes difficulties computationally, as two transfer matrices have to be determined. A better approach is to write the wave function in *every* region (r) as a linear combination of travelling waves, i.e.

$$\psi_r = A' \exp(ik_r z) + B' \exp(-ik_r z) \tag{2.164}$$

where the coefficients A' and B' are now allowed to be complex and the wave vector k is as before:

$$k = \sqrt{\frac{2m^*}{\hbar^2}(E - V_r)} \tag{2.165}$$

Forming the transfer matrix from this standard form for the wave function and applying the boundary condition of decaying exponentials at both ends of the well structure allows the method of solution as before. For now it is worthwhile confirming that ψ_r can take both the usual well- and barrier-state forms.

Consider regions where $E > V$:

$$\psi_r = A'(\cos k_r z + i \sin k_r z) + B'(\cos k_r z - i \sin k_r z) \tag{2.166}$$

$$\therefore \psi_r = (A' + B') \cos k_r z + i(A' - B') \sin k_r z \tag{2.167}$$

Since the coefficients are both complex numbers, they can be represented explicitly in terms of their real and imaginary components:

$$A' + B' = \Re(A' + B') + i\Im(A' + B') \tag{2.168}$$

and

$$A' - B' = \Re(A' - B') + i\Im(A' - B') \tag{2.169}$$

Substituting these into equation (2.167) and collecting real and imaginary components, then:

$$\begin{aligned} \psi_r = \ & [\Re(A' + B') \cos k_r z - \Im(A' - B') \sin k_r z] \\ & + i\,[\Im(A' + B') \cos k_r z + \Re(A' - B') \sin k_r z] \end{aligned} \tag{2.170}$$

Now eigenstates confined within potential wells are *stationary states*, which means that they have no time dependence and carry no current, i.e. their wave functions are real. Thus it would be expected that the solution would naturally yield such states, and in fact for the imaginary component of ψ_r to be zero, then:

$$\Im(A' + B') = 0 \quad \text{and} \quad \Re(A' - B') = 0 \tag{2.171}$$

which are simultaneously satisfied if:

$$B' = A'^* \tag{2.172}$$

This implies that in regions where the energy of the state of interest is greater than the potential, the solution will then give the coefficients as a complex conjugate pair in order to ensure a real wave function.

In regions where $E < V$, then:

$$k = i\sqrt{\frac{2m^*}{\hbar^2}(V_\mathrm{r} - E)} = i\kappa \tag{2.173}$$

where κ is the usual decay constant as defined in equation (2.130). Then the general form for a barrier state is generated, i.e.

$$\psi_\mathrm{r} = A' \exp(-\kappa z) + B' \exp(+\kappa z) \tag{2.174}$$

As mentioned above, application of the boundary conditions of $\psi \to 0$ as $z \to \pm\infty$ would force the choice that B' was zero and A' was real. In other regions where $E < V$, e.g. the step in Fig. 2.25 or the central barrier of Fig. 2.24, both A' and B' should naturally arise as real.

2.10 Addition of an electric field

As illustrated in Fig. 2.26, the effect of an electric field along the growth (z-)axis is to add a linear potential, which for an electron of charge $-e$ in the conduction band will be written as $-eFz$. For small electric fields, the effect of this electric field on the confined energy levels within a quantum well can be approximated by first-order perturbation theory.

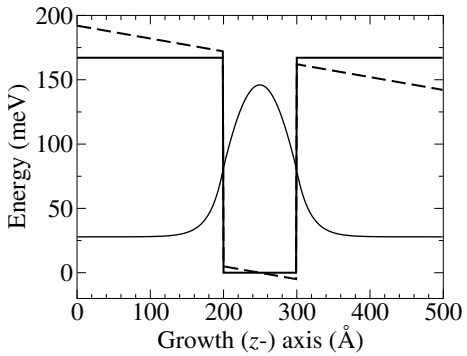


Figure 2.26: The ground state wave function of the zero-field potential (solid line) and the corresponding potential at a field of $10\,\mathrm{kV\,cm^{-1}}$ (dashed line)

If the perturbing potential is written as V' then the change in the ground state energy level is given, for example, by Schiff [18], p. 246:

$$\Delta E^{(1)} = \langle \psi_1 | V' | \psi_1 \rangle \tag{2.175}$$

with the electric field along the growth axis as the perturbation; this then translates to

$$\Delta E^{(1)} = \int_{-\infty}^{+\infty} \psi_1^*(z)(-eFz)\psi_1(z)\,\mathrm{d}z \tag{2.176}$$

The ground state of the *symmetric* quantum well of Fig. 2.26 is an even function, and hence the integrand of equation (2.176) is odd. Therefore, evaluation of the integral gives zero, i.e. to first order the addition of a *small* electric field has no effect on the *ground state* energy level. When applying the same logic to the first excited state, although ψ_2 is odd the integrand will still be an odd function, which again results in no change. There is currently much interest in the literature in breaking the symmetry of quantum wells in order to induce a first-order response in the ground state energy to an electric field [19, 20, 21].

Larger electric fields would require a more accurate evaluation of the perturbation as can be given by second-order perturbation theory (see Schiff [18], p. 247):

$$\Delta E^{(2)} = \sum_{m=2}^{\infty} \frac{|\langle \psi_m | V' | \psi_1 \rangle|^2}{E_m - E_1} \tag{2.177}$$

which in this case translates to:

$$\Delta E^{(2)} = \sum_{m=2}^{\infty} \frac{|\int_{-\infty}^{+\infty} \psi_m^*(z)(-eFz)\psi_1(z)\ \mathrm{d}z|^2}{E_m - E_1} \tag{2.178}$$

where the sum is over *all* excited states, including those with an energy $E_m > V$, i.e. the so-called 'continuum' states. As the electric field F does not have a z-dependence, then clearly the change in the energy to second-order $\Delta E^{(2)}$ is proportional to F^2. In addition, a charged particle prefers to move to areas of lower potential, i.e. the electron within the quantum well of Fig. 2.26 moves to the right-hand side of the well, thus lowering its total energy, so indeed $\Delta E^{(2)} \propto -F^2$. This suppression of the confined energy level by an electric field is known as the 'quantum-confined Stark effect' [22] and is commonly observed experimentally in heterostructures [20].

While such a perturbative treatise may be adequate for single quantum wells, its accuracy for complicated, perhaps multiple well systems and at high fields is questionable. In order to improve upon this, a full (non-perturbative) solution to the heterostructure and field potentials is required. Within the envelope function and effective mass approximations, the Schrödinger equation within a particular material (i.e. region of constant mass) and with an applied electric field of strength F is:

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2 \psi}{\partial z^2} + [V - eFz]\,\psi = E\psi \tag{2.179}$$

where V is taken as being a constant band-edge potential for the material under consideration. This new equation does not have the standard solution of linear combinations of trigonometric and exponential functions, and therefore a different approach is required. First, rearrange equation (2.179) to give:

$$\frac{\partial^2 \psi}{\partial z^2} - \frac{2m^*}{\hbar^2}\,[V - eFz - E]\,\psi = 0 \tag{2.180}$$

Then by making the substitutions:

$$\alpha = \frac{2m^*}{\hbar^2}\,[V - E] \tag{2.181}$$

and

$$\beta = \frac{2m^*}{\hbar^2}(-eF) \tag{2.182}$$

the Schrödinger equation becomes:

$$\frac{\partial^2\psi}{\partial z^2} - (\alpha + \beta z)\,\psi = 0 \tag{2.183}$$

Consider the further substitution:

$$z' = \frac{\alpha + \beta z}{\gamma} \tag{2.184}$$

where γ is an as yet unknown constant. Then, since z' is a first-order linear function of z:

$$\frac{\partial^2\psi}{\partial z^2} = \frac{\partial^2\psi}{\partial z'^2} \times \left(\frac{\partial z'}{\partial z}\right)^2 \tag{2.185}$$

i.e.

$$\frac{\partial^2\psi}{\partial z^2} = \frac{\beta^2}{\gamma^2}\frac{\partial^2\psi}{\partial z'^2} \tag{2.186}$$

Substituting into equation (2.183) then gives:

$$\frac{\beta^2}{\gamma^2}\frac{\partial^2\psi}{\partial z'^2} - \gamma z'\psi = 0 \tag{2.187}$$

If γ^3 is set equal to β^2, then the full transformation is:

$$z' = \left(\frac{2m^*}{\hbar^2}\right)^{\frac{1}{3}}\left[\frac{V-E}{(eF)^{\frac{2}{3}}} - (eF)^{\frac{1}{3}}z\right] \tag{2.188}$$

and the Schrödinger equation can be written as:

$$\frac{\partial^2\psi}{\partial z'^2} - z'\psi = 0 \tag{2.189}$$

The reason for such a procedure is that equation (2.189) has a standard solution which is a linear combination of Airy functions (see Abramowitz and Stegun [23], p. 446):

$$\psi(z') = A\ \mathrm{Ai}(z') + B\ \mathrm{Bi}(z') \tag{2.190}$$

The full solution proceeds by matching the wave functions at the heterojunctions according to the BenDaniel–Duke boundary conditions as before.

This is a standard procedure and many examples of its use can be found in the literature [24]. It is included along with the transfer matrix method of the previous sections for completeness. However, it can be appreciated that within these methods the study of a new quantum well system requires quite a large investment in developing new computational solutions. While a general computer code can be written to solve all semiconductor heterostructures, another one has to be written to solve all of these systems again with an electric field. The possibility of adding piezoelectric fields [25], which tilt the bands like an electric field, but with the additional possibility of being in the other direction, would require another computer code, and so on. While we continue in this vain for the remainder of this chapter, in the next chapter a numerical solution is introduced, which, although it occasionally has a few limitations, is still very powerful because of its versatility.

2.11 The infinite superlattice

Finite superlattices (or multiple quantum wells) can be treated with the transfer matrix technique, as described above for a double quantum well (a two-period finite superlattice). Such approaches yield a discrete set of states due to the confinement along the growth axis of the semiconductor multilayer. However, when many identical quantum wells are stacked within the same semiconductor layer the electrons and holes see an infinite periodic potential, similar to the Kronig–Penney model of bulk crystals. When this occurs, the electron and hole wave functions are no longer localised but are of infinite extent and equally likely to be in any of the quantum wells. They are said to occupy 'Bloch states'.

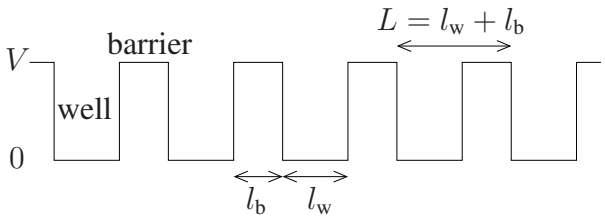


Figure 2.27: Band-edge potential profile of a superlattice of well and barrier widths l_w and l_b, respectively

As described in Section 2.9, and following Jaros [26], p. 220, the envelope function within the well region can be written as:

$$\psi_\mathrm{w} = A\exp\left(ik_\mathrm{w}z\right) + B\exp\left(-ik_\mathrm{w}z\right) \tag{2.191}$$

and in the barrier region as:

$$\psi_\mathrm{b} = C\exp\left(ik_\mathrm{b}z\right) + D\exp\left(-ik_\mathrm{b}z\right) \tag{2.192}$$

where:

$$k_\mathrm{w} = \sqrt{\frac{2m_\mathrm{w}^*}{\hbar^2}E}, \qquad k_\mathrm{b} = \sqrt{\frac{2m_\mathrm{b}^*}{\hbar^2}(E-V)} \tag{2.193}$$

and it is assumed (for now) that $E > V$, so that k_b is real. Clearly, and as mentioned above, if the superlattice is infinite, then a particle is equally likely to be found in any well, which means that its wave function must be periodic with the lattice, i.e.

$$\psi(z) = \psi(z+L) \tag{2.194}$$

For a travelling-wave state of the form $\exp\left(ikz\right)$, then:

$$\psi(z+L) = \exp\left[ik(z+L)\right] = \exp\left(ikz\right)\exp\left(ikL\right) \tag{2.195}$$

i.e.

$$\psi(z+L) = \psi(z)\exp\left(ikL\right) \tag{2.196}$$

By using this *periodic* form of the first BenDaniel–Duke boundary conditions (ψ continuous) at $z = L$, then:

$$\psi_{\mathrm{w}}(L) = \psi_{\mathrm{w}}(0)\exp(ikL) = \psi_{\mathrm{b}}(L) \tag{2.197}$$

Using the wave functions of equations (2.191) and (2.192), then:

$$(A+B)\exp(ikL) = C\exp(ik_{\mathrm{b}}L) + D\exp(-ik_{\mathrm{b}}L) \tag{2.198}$$

By employing the same periodicity condition in the second of the BenDaniel–Duke boundary conditions (ψ/m), also at $z = L$, then:

$$\frac{ik_{\mathrm{w}}}{m^*_{\mathrm{w}}}(A-B)\exp(ikL) = \frac{ik_{\mathrm{b}}}{m^*_{\mathrm{b}}}\left[C\exp(ik_{\mathrm{b}}L) - D\exp(-ik_{\mathrm{b}}L)\right] \tag{2.199}$$

Then using the BenDaniel–Duke boundary conditions at $z = l_{\mathrm{w}}$:

$$A\exp(ik_{\mathrm{w}}l_{\mathrm{w}}) + B\exp(-ik_{\mathrm{w}}l_{\mathrm{w}}) = C\exp(ik_{\mathrm{b}}l_{\mathrm{w}}) + D\exp(-ik_{\mathrm{b}}l_{\mathrm{w}}) \tag{2.200}$$

and

$$\frac{ik_{\mathrm{w}}}{m^*_{\mathrm{w}}}\left[A\exp(ik_{\mathrm{w}}l_{\mathrm{w}}) - B\exp(-ik_{\mathrm{w}}l_{\mathrm{w}})\right] = \frac{ik_{\mathrm{b}}}{m^*_{\mathrm{b}}}\left[C\exp(ik_{\mathrm{b}}l_{\mathrm{w}}) - D\exp(-ik_{\mathrm{b}}l_{\mathrm{w}})\right] \tag{2.201}$$

Equations (2.198)–(2.201) can be rewritten as:

$$\begin{pmatrix} \exp(ikL) & \exp(ikL) & -\exp(ik_{\mathrm{b}}L) & -\exp(-ik_{\mathrm{b}}L) \\ \frac{k_{\mathrm{w}}}{m^*_{\mathrm{w}}}\exp(ikL) & -\frac{k_{\mathrm{w}}}{m^*_{\mathrm{w}}}\exp(ikL) & -\frac{k_{\mathrm{b}}}{m^*_{\mathrm{b}}}\exp(ik_{\mathrm{b}}L) & \frac{k_{\mathrm{b}}}{m^*_{\mathrm{b}}}\exp(-ik_{\mathrm{b}}L) \\ \exp(ik_{\mathrm{w}}l_{\mathrm{w}}) & \exp(-ik_{\mathrm{w}}l_{\mathrm{w}}) & -\exp(ik_{\mathrm{b}}l_{\mathrm{w}}) & -\exp(-ik_{\mathrm{b}}l_{\mathrm{w}}) \\ \frac{k_{\mathrm{w}}}{m^*_{\mathrm{w}}}\exp(ik_{\mathrm{w}}l_{\mathrm{w}}) & -\frac{k_{\mathrm{w}}}{m^*_{\mathrm{w}}}\exp(-ik_{\mathrm{w}}l_{\mathrm{w}}) & -\frac{k_{\mathrm{b}}}{m^*_{\mathrm{b}}}\exp(ik_{\mathrm{b}}l_{\mathrm{w}}) & \frac{k_{\mathrm{b}}}{m^*_{\mathrm{b}}}\exp(-ik_{\mathrm{b}}l_{\mathrm{w}}) \end{pmatrix} \begin{pmatrix} A \\ B \\ C \\ D \end{pmatrix} = 0 \tag{2.202}$$

which implies that, for a solution other than the trivial $A = B = C = D = 0$, the determinant of the 4×4 matrix must be zero, i.e.

$$\begin{vmatrix} \exp(ikL) & \exp(ikL) & -\exp(ik_{\mathrm{b}}L) & -\exp(-ik_{\mathrm{b}}L) \\ \frac{k_{\mathrm{w}}}{m^*_{\mathrm{w}}}\exp(ikL) & -\frac{k_{\mathrm{w}}}{m^*_{\mathrm{w}}}\exp(ikL) & -\frac{k_{\mathrm{b}}}{m^*_{\mathrm{b}}}\exp(ik_{\mathrm{b}}L) & \frac{k_{\mathrm{b}}}{m^*_{\mathrm{b}}}\exp(-ik_{\mathrm{b}}L) \\ \exp(ik_{\mathrm{w}}l_{\mathrm{w}}) & \exp(-ik_{\mathrm{w}}l_{\mathrm{w}}) & -\exp(ik_{\mathrm{b}}l_{\mathrm{w}}) & -\exp(-ik_{\mathrm{b}}l_{\mathrm{w}}) \\ \frac{k_{\mathrm{w}}}{m^*_{\mathrm{w}}}\exp(ik_{\mathrm{w}}l_{\mathrm{w}}) & -\frac{k_{\mathrm{w}}}{m^*_{\mathrm{w}}}\exp(-ik_{\mathrm{w}}l_{\mathrm{w}}) & -\frac{k_{\mathrm{b}}}{m^*_{\mathrm{b}}}\exp(ik_{\mathrm{b}}l_{\mathrm{w}}) & \frac{k_{\mathrm{b}}}{m^*_{\mathrm{b}}}\exp(-ik_{\mathrm{b}}l_{\mathrm{w}}) \end{vmatrix} = 0 \tag{2.203}$$

Proceeding one step in evaluating the determinant would give:

$$
\begin{aligned}
0 = \; & \exp(ikL)\begin{vmatrix} -\frac{k_w}{m_w^*}\exp(ikL) & -\frac{k_b}{m_b^*}\exp(ik_bL) & \frac{k_b}{m_b^*}\exp(-ik_bL) \\ \exp(-ik_wl_w) & -\exp(ik_bl_w) & -\exp(-ik_bl_w) \\ -\frac{k_w}{m_w^*}\exp(-ik_wl_w) & -\frac{k_b}{m_b^*}\exp(ik_bl_w) & \frac{k_b}{m_b^*}\exp(-ik_bl_w) \end{vmatrix} \\
& -\exp(ikL)\begin{vmatrix} \frac{k_w}{m_w^*}\exp(ikL) & -\frac{k_b}{m_b^*}\exp(ik_bL) & \frac{k_b}{m_b^*}\exp(-ik_bL) \\ \exp(ik_wl_w) & -\exp(ik_bl_w) & -\exp(-ik_bl_w) \\ \frac{k_w}{m_w^*}\exp(ik_wl_w) & -\frac{k_b}{m_b^*}\exp(ik_bl_w) & \frac{k_b}{m_b^*}\exp(-ik_bl_w) \end{vmatrix} \\
& -\exp(ik_bL)\begin{vmatrix} \frac{k_w}{m_w^*}\exp(ikL) & -\frac{k_w}{m_w^*}\exp(ikL) & \frac{k_b}{m_b^*}\exp(-ik_bL) \\ \exp(ik_wl_w) & \exp(-ik_wl_w) & -\exp(-ik_bl_w) \\ \frac{k_w}{m_w^*}\exp(ik_wl_w) & -\frac{k_w}{m_w^*}\exp(-ik_wl_w) & \frac{k_b}{m_b^*}\exp(-ik_bl_w) \end{vmatrix} \\
& +\exp(-ik_bL)\begin{vmatrix} \frac{k_w}{m_w^*}\exp(ikL) & -\frac{k_w}{m_w^*}\exp(ikL) & -\frac{k_b}{m_b^*}\exp(-ik_bL) \\ \exp(ik_wl_w) & \exp(-ik_wl_w) & -\exp(ik_bl_w) \\ \frac{k_w}{m_w^*}\exp(ik_wl_w) & -\frac{k_w}{m_w^*}\exp(-ik_wl_w) & -\frac{k_b}{m_b^*}\exp(ik_bl_w) \end{vmatrix}
\end{aligned}
\tag{2.204}
$$

which the reader can see becomes very tedious indeed. Fortunately, use can be made of the constant-mass case of Jaros [26], generalised to account for $m_w^* \neq m_b^*$ as here, giving:

$$
\cos(k_wl_w)\cos(k_bl_b) - \sin(k_wl_w)\sin(k_bl_b)\left(\frac{(m_b^*)^2k_w^2 + (m_w^*)^2k_b^2}{2m_w^*m_b^*k_wk_b}\right) = \cos(kL) \tag{2.205}
$$

If $E < V$, then, following the arguments in Section 2.9, k_b becomes $i\kappa$, where:

$$
\kappa = \sqrt{\frac{2m_b^*}{\hbar^2}(V - E)} \tag{2.206}
$$

Substituting for k_b into equation (2.205), and using the following identities:

$$
\cos(i\kappa l_b) \equiv \cosh(\kappa l_b), \qquad \sin(i\kappa l_b) \equiv i\sinh(\kappa l_b) \tag{2.207}
$$

then:

$$
\cos(k_wl_w)\cosh(\kappa l_b) - \sin(k_wl_w)\sinh(\kappa l_b)\left(\frac{(m_b^*)^2k_w^2 - (m_w^*)^2\kappa^2}{2m_w^*m_b^*k_w\kappa}\right) = \cos(kL) \tag{2.208}
$$

The superlattice dispersion curves, i.e. the energy E of a particle as a function of its wave vector k, are obtained by solving equations (2.205) and (2.208). This is accomplished by using a similar method to that used in Section 2.7, i.e. equations (2.205) and (2.208) are rewritten in the form $f(E, k) = 0$ and solved for chosen values of k. In this case, however, there is no convenient way to identify the unique range of energies within which each solution lies.

Instead, a block search can be performed, in which the energy scale is divided up into small units of size δE, which is chosen to be smaller than adjacent solutions for E; typically $1\,\mu\text{eV}$

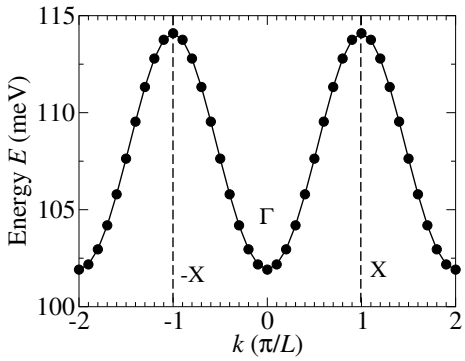


Figure 2.28: Lowest-energy solutions of the Kronig–Penney model of a $40\,$Å GaAs/$40\,$Å $Ga_{0.6}Al_{0.4}As$ superlattice, for a range of electron momenta along the growth (z-)axis

is a good choice. Each unit is searched for a sign change between $f(E,k)$ and $f(E+\delta E,k)$, indicating that a solution lies within that range. If a solution is detected within a cell, its precise value may be found using the methods in Section 3.1 or 3.2.

Figure 2.28 shows the lowest-energy solutions of a $40\,$Å GaAs/$40\,$Å $Ga_{0.6}Al_{0.4}As$ superlattice, for a range of k values. It can be seen that the energy E is a periodic function of k with period $2\pi/L$. The wave vector k was introduced earlier to describe the wave function as a travelling wave; clearly it can be identified as the momentum of a charge carrier along the growth (z-)axis of an *infinite* superlattice. Using Fig. 2.28, it can be seen that a definite minimum exists when the particle is at rest, i.e. $k=0$. As the carrier momentum increases, its energy increases too and reaches a maximum at $k=\pm\pi/L$. Thus the carrier within the superlattice occupies a continuous range of energies with a maxima and minima; this domain is the superlattice analogy to the energy bands of a crystal and is referred to as a *miniband*.

In addition the symmetry points of the E–k curves are labelled as in bulk crystals, where $k=0$ is called Γ and $k=\pi/L$ is referred to as X. This domain in k-space has become known as the *superlattice Brillouin zone* and, as in the bulk, is a convenient way to express the relationship between energy and momenta.

Just as single quantum wells can have more than one confined state, superlattices can have more than one miniband. Figure 2.29 displays the two lowest-energy solutions of the longer-period $60\,$Å GaAs/$60\,$Å $Ga_{0.6}Al_{0.4}As$ superlattice; note on this scale that the lowest-energy miniband looks almost flat. Clearly, the second miniband has the same periodicity as the first, although it should be noted that its minimum in energy occurs at the edge of the superlattice Brillouin zone and not at the centre. This effect is exploited in the 'interminiband laser' [27, 28, 29], which is a source of mid-infrared radiation.

The structural dependence of the miniband is illustrated in Fig. 2.30 for a series of superlattices with equal well and barrier widths. The two solid curves display the energy as a function of $l_{\rm w}$ at Γ and X. It can be seen that these energies fall with increasing well width, as does the miniband width, defined as follows:

$$\text{miniband width} = |E(\mathrm{X}) - E(\Gamma)| \tag{2.209}$$

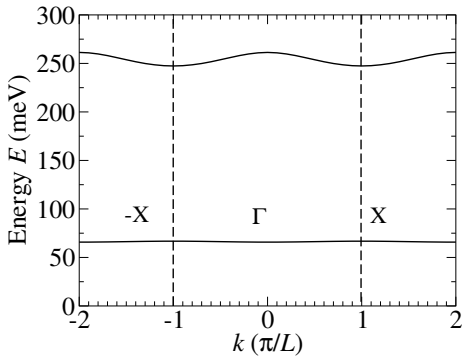


Figure 2.29: Two lowest-energy solutions of a 60 Å GaAs/60 Å $Ga_{0.6}Al_{0.4}As$ superlattice

where the modulus bars have been included to account for minibands such as the second, for which $E(\mathrm{X}) < E(\Gamma)$. Interestingly, the top and bottom of the miniband straddle the energy of the corresponding single quantum well (given by the dashed line). As the wells become more isolated and further apart, the miniband narrows and the energies tend towards that of the single quantum well. This represents the cross-over from a superlattice to a multiple quantum well—a collection of non-interacting identical wells.

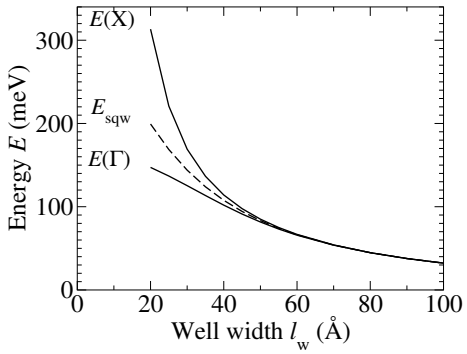


Figure 2.30: Energy at $k = 0$ (Γ) and $k = \pi/L$ (X), showing the lower and upper energy limits of the superlattice minibands, compared with the ground state energy of the single quantum well of the same width

As the miniband width changes with structure, the curvature of the E–k curve changes too. It is clear from Chapter 1 that the effective mass of a particle subject to an energy–momentum dependence of $E(\mathbf{k})$ can be derived from:

$$m^* = \frac{\hbar^2}{\partial^2 E/\partial \mathbf{k}^2} \tag{2.210}$$

Therefore, when an electron or hole is moving along the growth axis of a superlattice, the energy–momentum relationship is such that they appear to have a different effective mass from the bulk. Figure 2.31 plots this new effective mass at Γ for the range of structures in Fig. 2.30.

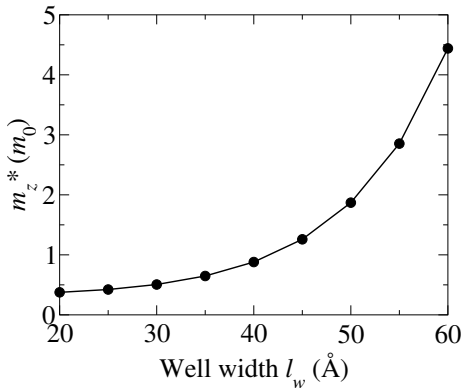


Figure 2.31: The new effective mass for an electron at Γ moving in the superlattice miniband

It can be seen from the figure that at very short periods the electron's miniband structure produces a new effective mass of $0.33m_0$, which is larger than the bulk effective mass for either the barrier or well materials ($0.067m_0$ and $0.10m_0$, respectively). Thus electrons at the edge of this miniband will have a much smaller response to an electric field than in bulk, just as from Newton's second law, $F = ma$, a larger m requires a larger force in order to produce the same acceleration.[1] In strict semiconductor language, this is quantified in terms of the carrier mobility, which is given by:

$$\mu = \frac{e\tau_{\text{coll}}}{m^*} \tag{2.211}$$

where τ_{coll} is the time between collisions (see [2], p. 601), the reciprocal of which is usually referred to as the total scattering rate for all mechanisms, via phonons, ionised impurities, other carriers, etc. Clearly, an increased effective mass within a miniband gives a reduced mobility μ for transport within that miniband, *provided* that all other scattering terms remain the same. This is an intuitively sound result, since the barriers within the superlattice would reasonably be expected to impede the transport of electrons when compared with a bulk material.

In this series of calculations, attention has been focused on superlattices with a simple single quantum well unit cell, and indeed with equal well and barrier widths. If these restrictions are lifted, then an almost infinite number of possible superlattices can be constructed, thus giving the opportunity to engineer, even design, the effective mass to suit the device.

[1] Previous editions of this book stated incorrectly that the effective mass is *smaller* in a superlattice than in bulk.

2.12 The single barrier

Hitherto, only semiconductor layered systems have been considered which form quantum wells, i.e. produce semiconductor layers which trap or confine the electrons along one axis. The 'opposite' of such structures also exist and are known as *barrier structures*. If a layer of a larger-band-gap material, e.g. $Ga_{1-x}Al_xAs$, is sandwiched between layers of a narrower-band-gap material, e.g. GaAs, then a potential barrier *can* result which repels carriers.

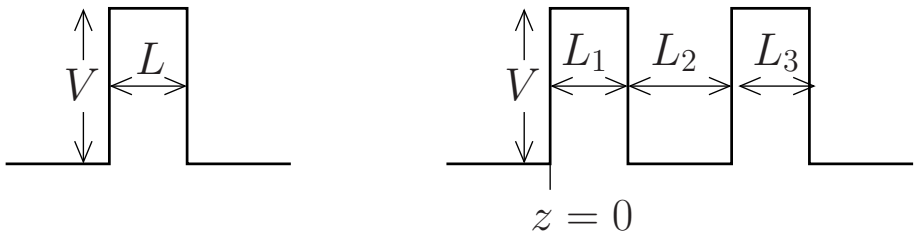


Figure 2.32: Single- and double-barrier structures

Figure 2.32 illustrates such a single-barrier structure. Electrons placed within this layered system simply collect in the GaAs outer (lower-energy) regions and there are no quantum-confined energy states. These systems do, however, exhibit quantum behaviour when an electric field is applied perpendicular to the layers (along the growth (z-)axis). Electrons (or holes) arising from doping are accelerated and impinge upon the barrier, and even when they have an energy E which is less than the potential energy height V of the barrier, they have a finite probability of passing through the barrier and appearing on the other side. This phenomenon is called *quantum mechanical tunnelling* or often just *tunnelling*. The bizarre nature of quantum mechanical tunnelling can be illustrated with a classical analogy: one would be pretty disturbed if, when kicking a football against a brick wall, it went straight through! Yet this is exactly what happens with *some* electrons and holes when they meet a potential barrier within a crystal. Tunnelling is an everyday phenomenon which occurs in a range of semiconductor devices, some of which appear in consumer electronics products (see, for example, Sze [30]).

One way of quantifying the proportion of electrons that tunnel through a barrier is in terms of the *transmission coefficient*, which is defined as the probability that any single electron impinging on a barrier structure will tunnel and contribute to the current flow through the barrier. Ferry has produced a comprehensive analysis of the transmission coefficient for a single-barrier structure (see [31], p. 60). Suffice here to quote Ferry's result ([31], equation (3.12)) for a constant effective mass across the structure—the transmission coefficient at an energy E for a barrier of width L and height V is given by:

$$T(E) = \frac{1}{1 + \left(\frac{k^2+\kappa^2}{2k\kappa}\right)^2 \sinh^2(\kappa L)}, \quad \text{for } E < V \tag{2.212}$$

where, as usual:

$$k = \frac{\sqrt{2m^*E}}{\hbar}, \qquad \kappa = \frac{\sqrt{2m^*(V-E)}}{\hbar} \tag{2.213}$$

For values of the carrier energy E greater than the barrier height V, $\kappa \to ik'$ (as in Section 2.11), and hence:

$$T(E) = \frac{1}{1 + \left(\frac{k^2-k'^2}{2kk'}\right)^2 \sin^2(k'L)}, \qquad \text{for } E > V \tag{2.214}$$

where

$$k = \frac{\sqrt{2m^*E}}{\hbar}, \qquad k' = \frac{\sqrt{2m^*(E-V)}}{\hbar} \tag{2.215}$$

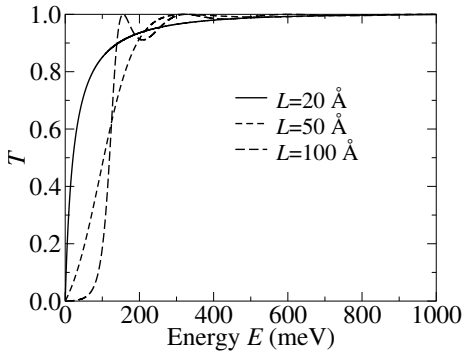


Figure 2.33: Transmission coefficient as a function of the energy through a single barrier for different barrier widths

The mathematics shows that, for $E > V$, the transmission coefficient would be expected to oscillate, with resonances when the sine term is zero. These occur when:

$$k'L = n\pi \tag{2.216}$$

which is implied when:

$$E = \frac{\hbar^2\pi^2n^2}{2m^*L^2} + V \tag{2.217}$$

The squared dependence implies that the resonances when $T = 1$ occur at larger and larger intervals in E, which can be clearly seen in Figs. 2.33 and 2.34. This is a remarkable result, since the energies at which resonances occur above the barrier have exactly the same form as the bound states within an infinite quantum well (equation (2.13)).

Figure 2.33 displays the transmission coefficient as a function of the energy E and for a range of barrier widths L. For this range of calculations, the barrier height V was fixed at $100\,\mathrm{meV}$, and below this energy the thinner the barrier, the higher the probability of tunnelling. For $E > V$, the situation is more complex due to the oscillatory nature of T. The

trend, however, as highlighted by the curves, is that the thicker the barrier, the closer the first resonance ($T = 1$) is to the top of the barrier. This can be understood from equation (2.217):

$$\lim_{L\to\infty} E^{n=1}_{\text{resonance}} = \lim_{L\to\infty} \frac{(\hbar\pi)^2}{2m^* L^2} + V = V \tag{2.218}$$

Conversely, for a fixed L and a variable V, the first resonance occurs at the same point above the barrier height; this is clearly illustrated in Fig. 2.34.

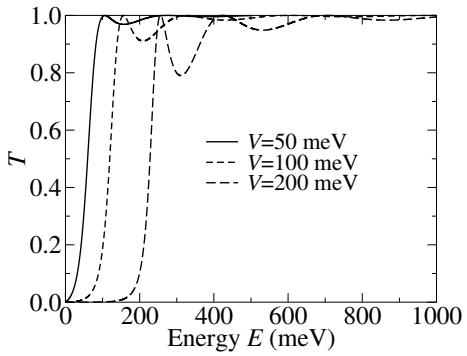


Figure 2.34: Transmission coefficient as a function of the energy through a single $100\,\text{Å}$ barrier for different barrier heights

2.13 The double barrier

If two barriers are placed a reasonably small distance apart (in the same crystal, perhaps a few nanometres) then the system is known as a *double barrier* (see Fig. 2.32), and has quite different transmission properties than the single barrier. Datta [32], p. 33, has deduced the transmission $T(E)$ dependence for the restricted case of symmetric barriers, while Ferry [31], p. 66, has considered asymmetric barriers. In this formalism, allowance will be made for differing barrier widths as well as the discontinuous change in the effective mass between well and barrier materials. Thus, with the aim of deducing the new $T(E)$, consider the solutions to Schrödinger's equation within each region for $E < V$:

$$\psi(z) = \begin{cases} A\exp(ikz) + B\exp(-ikz), & z < I_1 \\ C\exp(\kappa z) + D\exp(-\kappa z), & I_1 < z < I_2 \\ F\exp(ikz) + G\exp(-ikz), & I_2 < z < I_3 \\ H\exp(\kappa z) + J\exp(-\kappa z), & I_3 < z < I_4 \\ K\exp(ikz) + L\exp(-ikz), & I_4 < z \end{cases} \tag{2.219}$$

where k and κ have their usual forms as given in equation (2.213) and the positions of the interfaces have been labelled I_1, I_2, I_3, and I_4, respectively. Using the standard BenDaniel–Duke boundary conditions at each interface gives, for $z = I_1 = 0$:

$$A + B = C + D \tag{2.220}$$

$$\frac{1}{m_{\mathrm{w}}}(ikA - ikB) = \frac{1}{m_{\mathrm{b}}}(\kappa C - \kappa D) \tag{2.221}$$

for $z = I_2 = L_1$:

$$C\exp(\kappa I_2) + D\exp(-\kappa I_2) = F\exp(ikI_2) + G\exp(-ikI_2) \tag{2.222}$$

$$\frac{1}{m_{\mathrm{b}}}\left[\kappa C\exp(\kappa I_2) - \kappa D\exp(-\kappa I_2)\right] = \frac{1}{m_{\mathrm{w}}}\left[ikF\exp(ikI_2) - ikG\exp(-ikI_2)\right] \tag{2.223}$$

for $z = I_3 = L_1 + L_2$:

$$F\exp(ikI_3) + G\exp(-ikI_3) = H\exp(\kappa I_3) + J\exp(-\kappa I_3) \tag{2.224}$$

$$\frac{1}{m_{\mathrm{w}}}\left[ikF\exp(ikI_3) - ikG\exp(-ikI_3)\right] = \frac{1}{m_{\mathrm{b}}}\left[\kappa H\exp(\kappa I_3) - \kappa J\exp(-\kappa I_3)\right] \tag{2.225}$$

for $z = I_4 = L_1 + L_2 + L_3$:

$$H\exp(\kappa I_4) + J\exp(-\kappa I_4) = K\exp(ikI_4) + L\exp(-ikI_4) \tag{2.226}$$

$$\frac{1}{m_{\mathrm{b}}}\left[\kappa H\exp(\kappa I_4) - \kappa J\exp(-\kappa I_4)\right] = \frac{1}{m_{\mathrm{w}}}\left[ikK\exp(ikI_4) - ikL\exp(-ikI_4)\right] \tag{2.227}$$

The method of solution is the transfer matrix technique as before, writing the above equations in matrix form:

$$\mathsf{M}_1\begin{pmatrix}A\\B\end{pmatrix} = \mathsf{M}_2\begin{pmatrix}C\\D\end{pmatrix} \tag{2.228}$$

$$\mathsf{M}_3\begin{pmatrix}C\\D\end{pmatrix} = \mathsf{M}_4\begin{pmatrix}F\\G\end{pmatrix} \tag{2.229}$$

$$\mathsf{M}_5\begin{pmatrix}F\\G\end{pmatrix} = \mathsf{M}_6\begin{pmatrix}H\\J\end{pmatrix} \tag{2.230}$$

$$\mathsf{M}_7\begin{pmatrix}H\\J\end{pmatrix} = \mathsf{M}_8\begin{pmatrix}K\\L\end{pmatrix} \tag{2.231}$$

Then, as before, the coefficients of the outer regions can be linked by forming the transfer matrix, i.e.

$$\begin{pmatrix}A\\B\end{pmatrix} = \mathsf{M}_1^{-1}\mathsf{M}_2\mathsf{M}_3^{-1}\mathsf{M}_4\mathsf{M}_5^{-1}\mathsf{M}_6\mathsf{M}_7^{-1}\mathsf{M}_8\begin{pmatrix}K\\L\end{pmatrix} \tag{2.232}$$

Clearly, this 2×2 matrix equation still has four unknowns and cannot be solved—it is at this point that additional boundary conditions have to be imposed from physical intuition.

Whereas before, the standard boundary conditions, i.e. $\psi(z) \to 0$ as $z \to \pm\infty$, were used to solve for the confined states within quantum wells, in these barrier structures these are not appropriate since the travelling waves in the outer layers can have infinite extent. The standard procedure is to assume, quite correctly, that all of the charge carriers approach the double barrier from the same side, as would occur when as part of a biased device, as illustrated schematically in Fig. 2.35. Furthermore, if it is assumed that there are no further heterojunctions to the right of the structure, then no further reflections can occur and the wave function beyond the structure can only have a travelling-wave component moving to the right, i.e. the coefficient L must be zero.

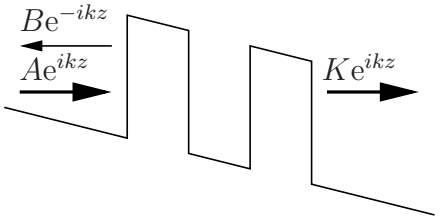


Figure 2.35: The wave function after imposition of the boundary conditions

Therefore, if the 2×2 matrix of equation (2.232) is written as M, then:

$$\begin{pmatrix} A \\ B \end{pmatrix} = \mathsf{M} \begin{pmatrix} K \\ 0 \end{pmatrix} \tag{2.233}$$

$$\therefore A = M_{11} K \tag{2.234}$$

and the ratio of transmitted to incident current, i.e. the transmission coefficient, is simply:

$$T(E) = \frac{K^* K}{A^* A} = \frac{1}{M_{11}^* M_{11}} \tag{2.235}$$

Figure 2.36 gives an example of the form of $T(E)$ for barriers of height $100\,\mathrm{meV}$ and width $100\,\text{Å}$ as a function of the distance L_2 between them. The effective masses were both taken to be equal to the bulk Γ valley electron mass of 0.067 m_0. It can be seen from the figure that the curves contain Dirac δ-functions at certain energies E below the potential barrier height V. This is quite unlike the single-barrier case. At these *resonance* energies, the double-barrier system appears transparent and has a transmission coefficient of 1. The wave functions of these states are localised between the barriers and are often referred to as *quasi-bound states* since they resemble the bound states of quantum well structures. Indeed, by comparing Fig. 2.36 with Fig. 2.17, the resonances can be seen to occur at exactly the same energies as the bound states in a finite well with the same width. However, they are not stationary states in that electrons or holes in such states will eventually scatter into the lower-energy states outside of the barriers.

The effect of an increasing barrier height V is shown in Fig. 2.37. It can be seen that, away from a resonance, an increasing barrier height leads, as would be expected, to a decrease

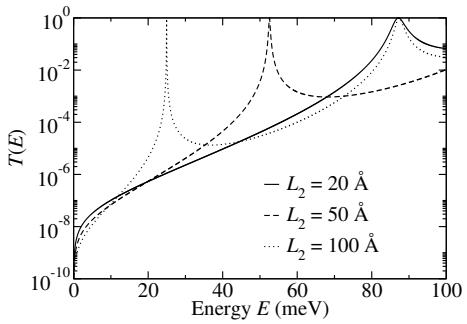


Figure 2.36: Transmission coefficient as a function of the energy through a double barrier of width $100\,\text{Å}$ and height $100\,\text{meV}$, separated by a distance L_2

in the transmission coefficient T. The classical explanation for this would be that it is harder for the electrons to tunnel through higher barriers. The *resonance energies* increase with increasing barrier height due to confinement effects, and the appearance of the second resonance at higher energies is a reflection on the existence of a second quasi-bound state.

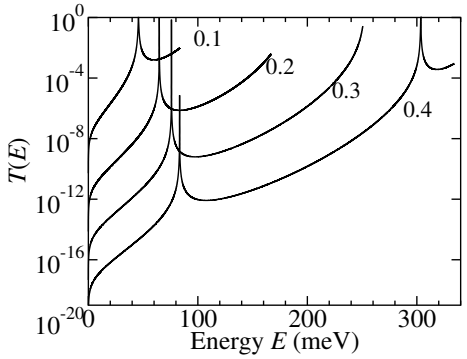


Figure 2.37: Transmission coefficient as a function of the energy through a $100\,\text{Å}$ $\text{Ga}_{1-x}\text{Al}_x\text{As}$/$50\,\text{Å}$ GaAs/$100\,\text{Å}$ $\text{Ga}_{1-x}\text{Al}_x\text{As}$ double barrier, for $x = 0.1, 0.2, 0.3, 0.4$

While the transmission coefficient represents a very important parameterisation of the properties of a double-barrier structure, it itself is not a measurable quantity. In fact, the properties of such two-terminal electronic devices are generally inferred (or summarised) from their current–voltage characteristics (I–V curve).

When an electric field is placed across such a double-barrier structure, any charge carriers present in the semiconductor, intrinsic or extrinsic, constitute a current which approaches the left-hand barrier. These charge carriers have a distribution of energy and momenta, often a Fermi–Dirac distribution. Those carriers that are of the same energy as the resonance are able to pass right through the double barrier without hindrance—a phenomenon that has become

known as *resonant tunnelling*. As the applied electric field (applied voltage) is increased, the number of carriers with the resonance energy increases and peaks as the Fermi level of the semiconductor to the left of the first barrier is brought into alignment. Therefore, the current gradually increases. At higher fields, the current falls away and a period of negative differential resistance ensues [33, 34]. Such a current–voltage characteristic has been exploited in high-frequency circuits, thus bringing the *resonant tunnelling diode* to prominence as a very useful electronic device [35].

There are various models of the current–voltage properties of different levels of complexity, the simplest of which would probably be to return to the idea that the current at any particular field would be equal to the number of carriers that tunnelled through the structure. This in turn would be the probability of a particular carrier tunnelling, multiplied by the number of carriers at that energy, i.e.

$$I \propto \int_{\text{band}} T(E) f^{\text{FD}}(E) \rho^{\text{3D}}(E)\, \mathrm{d}E \tag{2.236}$$

As the carriers approaching the barrier structure are in a bulk band, the integral is over their energies, and the Fermi–Dirac distribution function and density of states have the bulk (3D) forms.

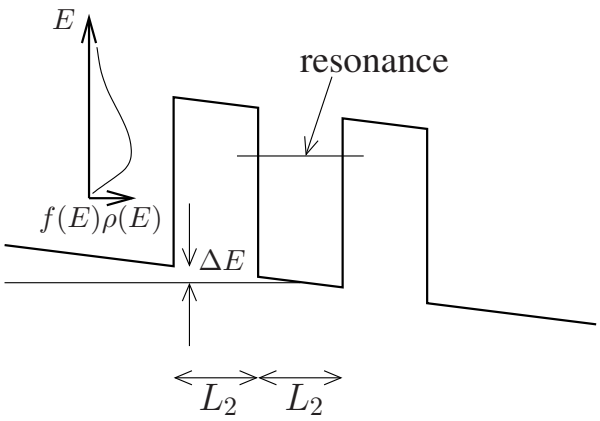


Figure 2.38: Simple model of current flow through a double-barrier structure

Figure 2.38 outlines this model. The electric field dependence is introduced via f^{FD} and ρ^{3D}. As the electric field is increased, the bottom of the bulk band is increased in energy relative to the centre of the double-barrier structure by an amount $\Delta E = eF(L_1 + L_2/2)$. Hence, the reference energy (the band minimum) will increase, and by using equation (2.49), the density of states at some energy E, measured from the band minimum at the centre of the well, would become:

$$\rho(E) = \frac{1}{2\pi^2} \left(\frac{2m^*}{\hbar^2} \right)^{\frac{3}{2}} (E - \Delta E)^{\frac{1}{2}}\, \Theta(E - \Delta E) \tag{2.237}$$

where the unit step function ensures that the energy of the carriers impinging on the left-hand barrier E is greater than ΔE. At the bottom of the band, $E = \Delta E$ and $\rho(E) = 0$. In addition, the Fermi energy of the bulk carriers increases by the same amount, i.e. from equation (2.68):

$$f^{\mathrm{FD}}(E) = \frac{1}{\exp\{[E - (E_{\mathrm{F}} + \Delta E)]/kT\} + 1} \tag{2.238}$$

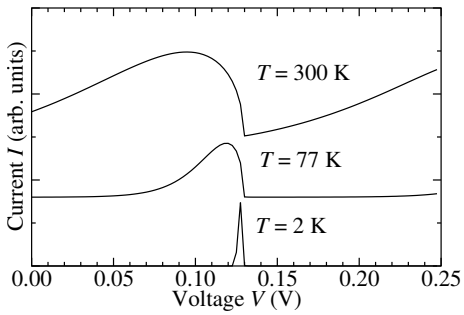


Figure 2.39: Current–voltage curve for the 100 Å $Ga_{0.8}Al_{0.2}As$/50 Å GaAs/100 Å $Ga_{0.8}Al_{0.2}As$ structure of earlier, obtained at different temperatures

Figure 2.39 illustrates the results obtained by using this model. The current is obtained in arbitrary units and the voltage V has been defined simply as the potential difference across the structure at a particular field, i.e. $V = F(L_1 + L_2 + L_3)$. The I–V curves have been plotted for several temperatures for the $x = 0.2$ structure of Fig. 2.37. The single resonance within this system shows itself as a single peak in the current. Clearly, the current peak broadens as the temperature increases, with this being a direct result of the broadening of the carrier distribution (the number of carriers in a given energy range), given by the product $f^{\mathrm{FD}}(E)\rho(E)$ appearing in the integral for the current, and illustrated schematically in Fig. 2.38. At low temperatures, the carriers occupy a small region of energy space around the band minima. As the field is increased, current only begins to flow when this narrow distribution is brought into line with the resonance energy, and hence there is a narrow current peak. As the temperature increases, the carrier distribution broadens, therefore there is a greater range of applied voltages that give some degree of alignment of the carriers with the resonance energy. The peak occurs when the peak of the distribution is aligned with the resonance energy. At voltages above this, the number of carriers available for tunnelling decreases, and hence the current also decreases.

The non-zero current at zero field for $T = 300$ K is a consequence of this simplistic model. It arises because even at zero field, there are a finite number of carriers in the broadened distribution which are aligned with the resonance energy and are therefore able to tunnel.

Figure 2.40 displays the results of the same simple model as a function of barrier width for the same structure of Fig. 2.39, but this time at the fixed temperature of 77 K. Again, as noted before, thinner barriers allow the electrons to tunnel more easily and hence give a higher current. An interesting physical point can be inferred from this data. The incomplete

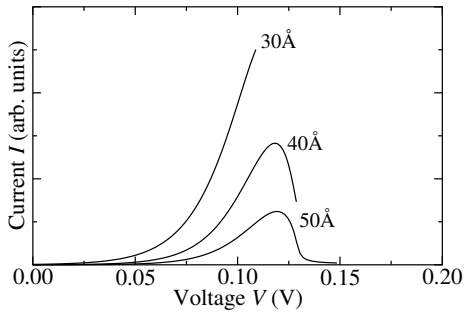


Figure 2.40: Current–voltage curves obtained at 77 K for a variety of barrier widths for the structure of Fig. 2.39

nature of the peaks in the 40 and 30 Å data is an indication that a significant fraction of the electrons are passing over the top of the barriers. While the current–voltage model advanced in this section can account for such a situation, the transmission coefficient versus energy data is terminated at the top of the barrier, and so only the fraction of electrons below this are included in the integration for the current in equation (2.236). It is left as an exercise for the reader to generalise the transmission coefficient theory to account for the case when $E > V$, and therefore improve the current–voltage model. This can be achieved by using a similar approach to that of Section 2.11.

The above discussion has been a simple introduction to the modelling of I–V curves for barrier structures, but nonetheless it shows some of the features of real devices. For a much more complete and in-depth study see, for example, Mizuta and Tanoue [35].

2.14 Extension to include electric field

An obvious improvement to the above model would be to account for the changes in the transmission coefficient as a function of the applied electric field. By using the substitution as before (equation (2.188)), i.e.

$$z' = \left(\frac{2m^*}{\hbar^2}\right)^{\frac{1}{3}} \left[\frac{V(z) - E}{(eF)^{\frac{2}{3}}} - (eF)^{\frac{1}{3}} z\right] \tag{2.239}$$

the solution in each region can then be written as a linear combination of Airy functions, just as for the general electric field case of equation (2.190), i.e.

$$\psi(z) = \begin{cases} A\,\mathrm{Ai}(z') + B\,\mathrm{Bi}(z'), & z < I_1 \\ C\,\mathrm{Ai}(z') + D\,\mathrm{Bi}(z'), & I_1 < z < I_2 \\ F\,\mathrm{Ai}(z') + G\,\mathrm{Bi}(z'), & I_2 < z < I_3 \\ H\,\mathrm{Ai}(z') + J\,\mathrm{Bi}(z'), & I_3 < z < I_4 \\ K\,\mathrm{Ai}(z') + L\,\mathrm{Bi}(z'), & I_4 < z \end{cases} \tag{2.240}$$

While Airy functions *can* be difficult to work with numerically, an immediate advantage over the zero-field case is that this solution is valid for both $E < V$ and $E > V$, and hence generalisation to this form produces two benefits. The method of solution is analogous to the zero-field case, in that application of the BenDaniel–Duke boundary conditions yields two equations for each interface, which in this case gives eight equations. The unknown coefficients, A and B, are linked to K and L as before by forming the transfer matrix, and are solved by imposition of a boundary condition, which is again a travelling wave in the direction of $+z$ to the right of the barrier structure. It is left to the interested reader to follow through such a derivation. A very general implementation for multiple barrier structures has been reported in the literature by Vatannia and Gildenblat [24].

2.15 Magnetic fields and Landau quantisation

If a magnetic field is applied externally to a *non-magnetic* semiconductor heterostructure then the constant effective mass Hamiltonian (familiar from equation (2.5)):

$$\mathcal{H} = -\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial z^2} + V(z) \tag{2.241}$$

which can be written:

$$\mathcal{H} = \frac{\mathcal{P}^2}{2m^*} + V(z) \tag{2.242}$$

becomes [36, 37, 38, 39]:

$$\mathcal{H} = \frac{1}{2m^*}\left(\mathcal{P} + e\mathbf{A}\right)^2 \mp \frac{1}{2}g^*(z)\mu_{\mathrm{B}}B + V(z) \tag{2.243}$$

where the kinetic energy operator becomes modified by the magnetic field vector potential $\mathbf{A}$, and the second term produces a splitting, known as the 'gyromagnetic spin splitting' between the spin-up ($-$ sign) and spin-down ($+$ sign) electrons. g^* is known as the 'Landé factor' and is really a function of z as it depends on the material; however, it is generally assumed to be constant and approximately 2 for conduction-band electrons. μ_{B} is the Bohr magneton and B is the magnitude of the magnetic flux density which is assumed aligned along the growth (z-)axis.

Although the heterostructure potential $V(z)$ remains one-dimensional, the vector potential means that the wave functions are not necessarily one-dimensional, so the Schrödinger equation must be written:

$$\left[\frac{1}{2m^*}\left(\mathcal{P} + e\mathbf{A}\right)^2 \mp \frac{1}{2}g^*(z)\mu_{\mathrm{B}}B + V(z)\right]\psi(x,y,z) = E\psi(x,y,z) \tag{2.244}$$

The magnetic field produces a parabolic potential along one of the in-plane axes (the x-axis, say), leaving the particle free to move (with a wave vector k_y, say) along the other axis. The standard approach is to employ the Landau gauge $\mathbf{A} = Bx\mathbf{e}_y$; then, following the notation of Savić *et al.* [40], the wave function can be written in the separable form $\psi(x,y,z) = \psi_x(x)\psi_y(y)\psi_z(z)$, i.e.

$$\psi(x,y,z) = \frac{1}{\sqrt{L_y}}\psi_j(x - X_{k_y})\mathrm{e}^{ik_y y}\psi(z) \tag{2.245}$$

where L_y is a normalisation constant (the length of the structure along the y-axis, $X_{k_y} = -k_y l_{\mathrm{B}}^2$, where $l_{\mathrm{B}} = \sqrt{\hbar/eB}$ is the Landau length), j is an index over the $\psi_j(x - X_{k_y})$ harmonic oscillator solutions of the parabolic potential and $\psi(z)$ is the usual one-dimensional envelope function of the heterostructure potential without the magnetic field. The harmonic oscillator solutions will be generated numerically in Section 3.10; however, they can also be expressed analytically as:

$$\psi_j(x - X_{k_y}) = \frac{1}{\pi^{\frac{1}{4}}\sqrt{2^j j! l_{\mathrm{B}}}} \exp\left(-\frac{(x - X_{k_y})^2}{2l_{\mathrm{B}}^2}\right) H_j\left(\frac{x - X_{k_y}}{l_{\mathrm{B}}}\right) \tag{2.246}$$

where H_j is the jth Hermite polynomial (see Liboff [41], p. 200, for a detailed description of the analytical solutions of the harmonic oscillator.

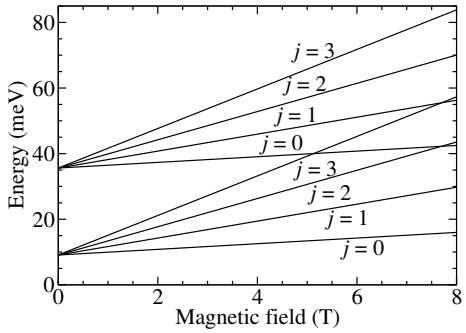


Figure 2.41: The magnetic field induced Landau levels of a $200\,$Å GaAs quantum well surrounded by $Ga_{1-x}Al_xAs$ barriers of height $100\,\mathrm{meV}$ (same structure as Fig. 2.18), with constant effective mass of $0.067m_0$. Note that only the first four Landau levels associated with each quantum well subband are shown for clarity

Taking g^* as a constant equal to 2 and the Bohr magneton as $9.274 \times 10^{-24}\,\mathrm{J\,T^{-1}}$, it can be seen that even in a relatively high magnetic field of $10\,\mathrm{T}$, the difference in energy between the spin-up and spin-down electrons is:

$$g^*\mu_{\mathrm{B}}B \approx 2 \times 9.274 \times 10^{-24}\,\mathrm{J\,T^{-1}} \times 10\,\mathrm{T} \approx 1\,\mathrm{meV} \tag{2.247}$$

which on the scales of typical quantum well systems and typical carrier densities is relatively small and can generally be ignored. Thus the total energy can be considered to be composed of just two components: the usual energy eigenvalue E_n associated with the nth state of the quantum well without a magnetic field and the harmonic oscillator type energy associated with the in-plane cyclotron motion $(j + \frac{1}{2})\hbar\omega_{\mathrm{c}}$, i.e.

$$E_{n,j} = E_n + \left(j + \frac{1}{2}\right)\hbar\omega_{\mathrm{c}} \tag{2.248}$$

where the 'cyclotron frequency' is the usual one from bulk semiconductors [2, 42, 43]:

$$\omega_{\mathrm{c}} = \frac{eB}{m^*} \tag{2.249}$$

Figure 2.41 shows the Landau level splitting induced by the application of an external magnetic field on a $200\,\text{Å}$ GaAs quantum well surrounded by $Ga_{1-x}Al_xAs$ barriers, assuming a constant electron effective mass of $0.067m_0$. The figure illustrates the lowest four Landau levels for each of the two lowest-energy confined states within the quantum well. A more accurate calculation could include the variation in effective mass of the carrier between the various layers; this would be done by calculating the weighted mean or expectation value of the effective mass according to the probability of finding the carrier in each of the semiconductor layers.

2.16 In summary

The consideration of simple layered semiconductor heterostructures with analytical forms for the solutions to Schrödinger's equation has allowed exploration and discovery of the properties of two-dimensional systems. Such models are invaluable and allow a whole range of physical observables of experimental layer structures and electronic devices to be explained. However, implementing these methods computationally *can* be tedious, in that a computer program, written to calculate the energy levels of a single quantum well, has to be rewritten for a double quantum well. In addition, a different program would be needed to solve a triangular well, and some potential profiles, which can be fabricated, such as diffused quantum wells, have no analytical solution at all. While such continuously varying structures can be expressed with flat step potentials for each monolayer [44], the treatment of an electric field in this manner is questionable.

In the next chapter, a simple, but very general *numerical* solution to Schrödinger's equation will be derived, which will overcome all of these difficulties.

Exercises

All necessary material parameters are given in Appendix A. Some of these exercises require the use of the QWWAD simulation tools. A brief introduction is given in Appendix B, along with links to download the programs and access further support.

(1). Consider a $10\,\text{nm}$ thick layer of semiconductor material, surrounded by impenetrable barriers.

   (a) Use the effective mass approximation to estimate the energy of the first subband minimum for electrons when the semiconductor layer consists of (i) GaAs, (ii) AlAs and (iii) $Cd_{0.85}Mn_{0.15}Te$.
   (b) Repeat this calculation to estimate the first heavy-hole subband minimum in each material.
   (c) Determine the energy difference between the lowest electron and hole states in each material.

(d) Photoluminescence techniques can be used experimentally to measure the difference between electron and hole states in semiconductor systems. Assuming that the measurement system has a $1\,\mathrm{meV}$ sensitivity, estimate the maximum thickness semiconductor layer in which the lowest-energy subbands can be observed for each material.

(2). (a) Find an expression for the *three-dimensional* density of states for an infinite quantum well by dividing equation (2.66) by the width of the well.

(b) Show that this result agrees with the bulk density of states in equation (2.49) when the energy is equal to the subband minima in the well.

(c) Find an expression for the difference $\Delta\rho^{3\mathrm{D}}$ between the 3D density of states in an infinite quantum well and a bulk semiconductor at an energy ΔE above the bottom of subband j. Show that $\Delta\rho^{3\mathrm{D}}$ vanishes as $j \to \infty$.

(d) Find the maximum value of $\Delta\rho^{3\mathrm{D}}$ by letting ΔE equal the separation between a pair of subbands. Hence, show that $\Delta\rho^{3\mathrm{D}}$ vanishes as $l_{\mathrm{w}} \to \infty$.

(3). The dispersion relation for a photon is $E = \hbar ck$. Find an expression for the density of photon states in a 3D space.

(4). Maxwell–Boltzmann statistics can be used to approximate the distribution of carriers in a subband, where $f^{\mathrm{MB}} = \exp[(E - E_{\mathrm{F}})/kT]$.

(a) Find an expression for the population of a subband using this approximation.

(b) Find an expression for the quasi-Fermi energy of a subband with a given population.

(c) Find the errors in the calculation of (a) and (b) when compared with the use of Fermi–Dirac statistics.

(d) Hence, determine the conditions under which Maxwell–Boltzmann statistics are a good approximation to Fermi–Dirac statistics.

(5). Use the `qwwad_ef_square_well` computer program to investigate the subband energies in a layer of GaAs surrounded by thick $\mathrm{Ga}_{0.85}\mathrm{Al}_{0.15}\mathrm{As}$ barriers.

(a) Assuming parabolic dispersion, determine the minimum well widths for which two, three or four electron subbands are present in the well.

(b) Determine the difference in energy between the lowest electron and hole subbands as a function of well width. Compare the results to those found for an infinite GaAs well in question (1).

(6). Consider a system in which an electron is confined in an extremely deep 1 Å wide well, such that it can be approximated using the potential $V(z) = -V_0 l_{\mathrm{w}} \delta(z)$, where $\delta(z)$ is the Dirac delta function.

(a) Use the `qwwad_ef_square_well` program to investigate systems of this type. How many bound states exist in the well? What is the binding energy (i.e. the difference between the energy of the state(s) and the top of the well)? How does this energy change with respect to V_0?

(b) Derive the wave function for the electron by noting that it must decay exponentially on either side of the narrow well. The well can be assumed to be of negligible width.

(c) Determine the change in the first derivative of the wave function on either side of the well by integrating the Schrödinger equation over a small region of width 2ϵ around the well:

$$\left.\frac{\mathrm{d}\psi}{\mathrm{d}z}\right|_{\text{right}} - \left.\frac{\mathrm{d}\psi}{\mathrm{d}z}\right|_{\text{left}} = \frac{2m^*}{\hbar^2}\int_{-\epsilon}^{+\epsilon}[V_0 l_{\text{w}}\delta(z) - E]\psi(z)\,\mathrm{d}z \tag{2.250}$$

(d) Hence, find the wave vector and energy of the bound state in the well as a function of V_0.

(7). In Section 2.11 it was shown that the effective mass in a superlattice converges to a fixed value when the period length becomes very small. Use the `qwwad_ef_superlattice` program to further investigate the behaviour of an electron at Γ in superlattices with very thin layers.

(a) Consider a structure with equal barrier and well widths. Determine, through numerical simulations, the relationship between the energy of the electron and the barrier potential.

(b) Similarly, show that the effective mass in the superlattice is independent of barrier potential, and determine its relationship to the effective masses in the well and barrier materials.

(c) Find more general expressions for the energy and effective mass of the electron when the ratio of $l_{\text{w}} : l_{\text{b}}$ is allowed to change.

(d) Determine the energy and effective masses in the limiting cases, when the well layers are much larger than the barrier layers and vice versa. Discuss whether these results are intuitively correct.

(e) Starting from equation (2.208), show that the expression for energy is correct when the layer widths tend to zero. (Hint: expand the trigonometric and hyperbolic functions in terms of Maclaurin series.)

(8). Experimentally, there is always a degree of imprecision in the growth of layers in a heterostructure. Use the `qwwad_tx_double_barrier_iv` program to investigate the effects of errors in layer thickness and alloy composition on the performance of a resonant-tunnelling diode (RTD). Example topics for investigation could include (among others) the energies of resonant states, the transmission linewidth (i.e. full width at half the maximum of the transmission peak), and the peak transmission probability. The study could be extended to a range of different material systems, well widths and temperatures. Use this study to determine an indicative set of tolerances (i.e. maximum tolerable errors) for the growth of RTD structures.

References

[1] J. W. Leech, *Classical Mechanics*, Chapman & Hall, London, Second edition, 1965.

[2] N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Saunders College Publishing, Philadelphia, 1976.

[3] J. A. López-Villanueva, F. Gámiz, I. Melchor, and J. A. Jiménez-Tejada, 'Density of states of a two-dimensional electron gas including nonparabolicity', *J. Appl. Phys.*, **75**(8):4267–4269, 1994.

[4] A. N. Chakravarti and B. R. Nag, 'Generalized Einstein relation for degenerate semiconductors having non-parabolic energy bands', *Int. J. Electron.*, **37**(2):281–284, 1974.

[5] M. P. Hasselbeck and P. M. Enders, 'Electron–electron interactions in the nonparabolic conduction band of narrow-gap semiconductors', *Phys. Rev. B*, **57**(16):9674–9681, 1998.

[6] G. Bastard, *Wave Mechanics Applied to Semiconductor Heterostructures*, Monographies de physique. Halsted Press, New York, 1988.

[7] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, London, Sixth edition, 2000.

[8] M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, and F. Rossi, *GNU Scientific Library Reference Manual.*, Network Theory Limited, Bristol, Third (v1.12) edition, 2009.

[9] R. M. Eisberg, *Fundamentals of Modern Physics*, John Wiley & Sons, Inc., New York, 1961.

[10] M. J. Kelly, *Low Dimensional Semiconductors: Materials, Physics, Technology, Devices*, Clarendon Press, Oxford, 1995.

[11] Z. Ikonić, V. Milanović, and D. Tjapkin, 'Bound–free intraband absorption in GaAs–$Al_xGa_{1-x}As$ semiconductor quantum wells', *Appl. Phys. Lett.*, **54**(3):247–249, 1989.

[12] G. Bastard, 'Superlattice band structure in the envelope-function approximation', *Phys. Rev. B*, **24**(10):5693–5697, 1981.

[13] G. T. Einevoll and L. J. Sham, 'Boundary conditions for envelope functions at interfaces between dissimilar materials', *Phys. Rev. B*, **49**(15):10533–10543, 1994.

[14] I. Galbraith and G. Duggan, 'Envelope-function matching conditions for GaAs/(Al,Ga)As heterojunctions', *Phys. Rev. B*, **38**(14):10057–10059, 1988.

[15] J. W. Conley, C. B. Duke, G. D. Mahan, and J. J. Tiemann, 'Electron tunneling in metal–semiconductor barriers', *Phys. Rev.*, **150**(2):466–469, 1966.

[16] D. J. BenDaniel and C. B. Duke, 'Space-charge effects on electron tunneling', *Phys. Rev.*, **152**(2):683–692, 1966.

[17] W. E. Hagston, P. Harrison, T. Piorek, and T. Stirner, 'Boundary conditions on current carrying states and the implications for observation of Bloch oscillations', *Superlatt. Microstruct.*, **15**:199–202, 1994.

[18] L. I. Schiff, *Quantum Mechanics*, McGraw-Hill, New York, 1968.

[19] J. Khurgin, 'Novel configuration of self-electro-optic effect device based on asymmetric quantum wells', *Appl. Phys. Lett.*, **53**:779, 1988.

[20] M. Morita, K. Goto, and T. Suzuki, 'Quantum-confined Stark effect in stepped-potential quantum wells', *Jpn. J. Appl. Phys.*, **29**(9A):L1663, 1990.

[21] N. Susa and T. Nakahara, 'Large blue shifts induced by Stark effect in asymmetric coupled quantum well', *Electr. Lett.*, **28**:941, 1992.

[22] D. A. B. Miller, D. S. Chemla, T. C. Damen, A. C. Gossard, W. Wiegmann, T. H. Wood, and C. A. Burrus, 'Band-edge electroabsorption in quantum well structures: The quantum-confined Stark effect', *Phys. Rev. Lett.*, **53**(22):2173–2176, 1984.

[23] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover Publications, New York, 1965.

[24] S. Vatannia and G. Gildenblat, 'Airy's function implementation of the transfer-matrix method for resonant tunneling in variably spaced finite superlattices', *IEEE J. Quant. Electr.*, **32**:1093, 1996.

[25] M. P. Halsall, J. E. Nicholls, J. J. Davies, B. Cockayne, and P. J. Wright, 'CdS/CdSe intrinsic Stark superlattices', *J. Appl. Phys.*, **71**(2):907–915, 1992.

[26] M. Jaros, *Physics and Applications of Semiconductor Microstructures*, Clarendon Press, Oxford, 1989.

[27] L. Friedman, R. Soref, and G. Sun, 'Quantum parallel laser: a unipolar superlattice interminiband laser', *IEEE Photonics Technol. Lett.*, **9**(5):593–595, 1997.

[28] G. Scamarcio, F. Capasso, J. Faist, C. Sirtori, D. L. Sivco, A. L. Hutchinson, and A. Y. Cho, 'Tunable interminiband infrared emission in superlattice electron transport', *Appl. Phys. Lett.*, **70**:1796–1798, 1997.

[29] G. Scamarcio, F. Capasso, C. Sirtori, J. Faist, A. L. Hutchinson, D. L. Sivco, and A. Y. Cho, 'High-power infrared (8-micrometer wavelength) superlattice lasers', *Science*, **276**(5313):773–776, 1997.

[30] S. M. Sze, *Physics of Semiconductor Devices*, John Wiley & Sons, Inc., New York, Second edition, 1981.

[31] D. K. Ferry, *Quantum Mechanics: An Introduction for Device Physicists and Electrical Engineers*, IOP Publishing, London, 1995.

[32] S. Datta, *Quantum Phenomena*, vol. VIII of *Modular Series on Solid State Devices*, Addison-Wesley, New York, 1989.

[33] S. Luyri, 'Hot-electron transistors,' in *High Speed Semiconductor Devices*, S. M. Sze, Ed., p. 399. Wiley-Interscience, New York, 1990.

[34] J. P. Sun, G. Haddad, P. Mazumder, and J. Schulman, 'Resonant tunneling diodes: models and properties', *Proc. IEEE*, **86**(4):641–660, 1998.

[35] H. Mizuta and T. Tanoue, *The Physics and Applications of Resonant Tunnelling Diodes*, Cambridge University Press, Cambridge, 1995.

[36] J. K. Furdyna, 'Diluted magnetic semiconductors', *J. Appl. Phys.*, **64**(4):R29–R64, 1988.

[37] J. A. Gaj, 'Magnetooptical properties of large-gap diluted magnetic semiconductors,' in *Diluted Magnetic Semiconductors*, J. K. Furdyna and J. Kossut, Eds., vol. 25. Academic Press, Boston, 1988.

[38] N. Malkova and U. Ekenberg, 'Spin properties of quantum wells with magnetic barriers. I. A k.p analysis for structures with normal band ordering', *Phys. Rev.*, **66**:155324, 2002.

[39] T. Wenckebach, *Essentials of Semiconductor Physics*, John Wiley & Sons, Ltd., Chichester, 1999.

[40] I. Savić, V. Milanović, Z. Ikonić, D. Indjin, V. Jovanović, and P. Harrison, 'Dilute magnetic semiconductor quantum-well structures for magnetic field tunable far-infrared/Terahertz absorption', *IEEE J. Quant. Electr.*, **40**:1614–1621, 2004.

[41] R. L. Liboff, *Introductory Quantum Mechanics*, Addison–Wesley, San Francisco, Fourth edition, 2003.

[42] J. S. Blakemore, *Solid State Physics*, Cambridge University Press, Cambridge, Second edition, 1985.

[43] R. Turton, *The Physics of Solids*, Oxford University Press, Oxford, 2000.

[44] A. R. Sugg and J.-P. C. Leburton, 'Modeling of modulation-doped multiple-quantum-well structures in applied electric fields using the transfer-matrix technique', *IEEE J. Quant. Electr.*, **27**:224, 1991.

3

Numerical solutions

Chapter 2 focused principally upon finding *analytical* solutions for the behaviour of carriers in heterostructures. In some cases, such as the infinite quantum well discussed in Section 2.1, it was possible to find *closed-form solutions* for the energies and wave functions. In other words, the solutions were expressed in terms of well-understood, simple mathematical functions ($\sin$, $\cos$, etc.), allowing the energies and wave functions to be calculated directly.

In general, however, closed-form solutions cannot be found for anything beyond the simplest systems. Indeed, even the single finite well system in Section 2.7 has no closed-form solution, and instead the solutions were found iteratively using numerical methods. This chapter begins by describing a few of these root-finding algorithms that allow solutions of a complicated mathematical function to be found.

Modern semiconductor devices can contain exceedingly complex heterostructures. One example is the quantum cascade laser [1], which contains hundreds of layers of semiconductor materials with a range of different thicknesses, with the band structure being distorted by large internal and external electric fields. In such cases, it may be impossible to even determine an analytical expression to be solved for the states in the system. Instead, discrete mathematical techniques can be used to create a *sampled approximation of the system* which can be solved computationally. Although these methods do not provide the same direct insight into the behaviour of systems as the analytical solutions in Chapter 2, they are generally applicable and very flexible, and the same techniques can be used to describe a vast range of structures. This chapter focuses on a range of such methods and their application to more complex heterostructures.

3.1 Bisection root-finding

The bisection root-finding technique is used to find solutions of an equation of the form:

$$f(x) = 0 \tag{3.1}$$

when the solutions to be found, x_n, are each known to lie within a particular range of values a_n and b_n.

A suitable example is the finite quantum well, whose solutions are given by equation (2.101). Rearranging the expression for the even-parity states into the form required by the bisection method gives:

$$f(v) = \sqrt{u_0^2 - v^2} - v\tan v = 0 \tag{3.2}$$

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

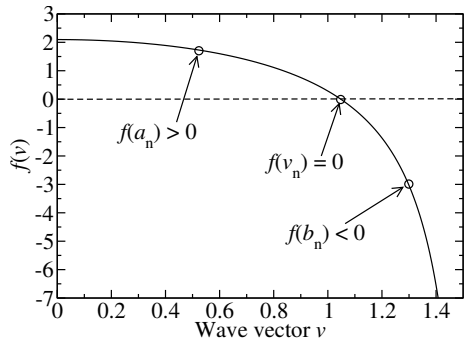


Figure 3.1: Matching equation for a 200 Å finite GaAs quantum well with $100\,\mathrm{meV}$ barriers. Equation (3.2) is plotted over a range of normalised wave vectors. Highlighted points correspond to a solution v_n, a wave vector smaller than the solution a_n and a wave vector larger than the solution b_n

and for the odd states:

$$f(v) = \sqrt{u_0^2 - v^2} + v\cot v = 0 \tag{3.3}$$

The values of this expression for the ground state are plotted over a range of wave vectors in Fig. 3.1. The plot may now be searched visually for the ground state wave vector by finding the point at which $f(v) = 0$, i.e. the point at which the function crosses the horizontal axis in the plot.

Furthermore, note that the function is positive for wave vectors below the solution and negative for wave vectors above the solution. In other words, a solution can be found between points where the sign of a function changes. This is the basis of the bisection search method.

The bisection algorithm proceeds as follows:

(1). *Bisect the limits* to find the midpoint $x = (a + b)/2$. This represents the best estimate of the solution of the function so far.

(2). *Compute the signs* of the function at x, a and b.

(3). *Revise the limits* to restrict attention to the range of values where a sign change occurs.

- If $f(x)$ has the same sign as $f(a)$, then the solution cannot lie between a and x. Replace a with x.
- If $f(x)$ has the same sign as $f(b)$, then the solution cannot lie between b and x. Replace b with x.

(4). *Repeat* steps 1–3 until the solution has been found to the desired precision $\delta x = b - a$.

For the finite well example, the upper and lower bounds can be specified for v in the ground state using equation (2.105), such that $a = 0$ and $b = \pi/2$. Table 3.1 shows the progression of the bisection search algorithm. After 11 steps, it converges to the solution $v = 1.265 \pm 0.002$, which corresponds to an energy:

$$E = \frac{2\hbar^2 v^2}{l_{\mathrm{w}}^2 m_w^*} = 9.12\,\mathrm{meV} \tag{3.4}$$

Table 3.1 Bisection search for the normalised wave vector of the ground state in a 200 Å finite quantum well with 100 meV barriers. A precision of $\delta v = 0.002$ was used

Iteration	a	b	$f(a)$	$f(b)$	v	δv
1	0.00000000	1.57079633	4.19349324	$-\infty$	0.78539816	1.57079633
2	0.78539816	1.57079633	3.33389004	$-\infty$	1.17809724	0.78539816
3	1.17809724	1.57079633	1.18043003	$-\infty$	1.37444678	0.39269908
4	1.17809724	1.37444678	1.18043003	−2.94795725	1.27627201	0.19634954
5	1.17809724	1.27627201	1.18043003	−0.21274425	1.22718463	0.09817477
6	1.22718463	1.27627201	0.58016189	−0.21274425	1.25172832	0.04908739
7	1.25172832	1.27627201	0.21328386	−0.21274425	1.26400017	0.02454369
8	1.26400017	1.27627201	0.00854279	−0.21274425	1.27013609	0.01227185
9	1.26400017	1.27013609	0.00854279	−0.09990688	1.26706813	0.00613592
10	1.26400017	1.26706813	0.00854279	−0.04515007	1.26553415	0.00306796
11	1.26400017	1.26553415	0.00854279	−0.01817264	1.26476716	0.00153398

as shown in Fig. 2.17.

The convergence of the bisection search is predictable, with the error δx being reduced by half at each iteration. As such, if the initial separation between the search limits is $\delta x_0 = b_0 - a_0$, then the error after N iterations is:

$$\delta x = \frac{\delta x_0}{2^N} \tag{3.5}$$

This can be rearranged to determine the number of iterations required to obtain a precision of δx, given a starting precision of δx_0:

$$2^N = \frac{\delta x_0}{\delta x} \tag{3.6}$$

Now, taking the natural logarithm of both sides and dividing by $\ln 2$:

$$N = \frac{\ln(\delta x_0) - \ln(\delta x)}{\ln 2} \tag{3.7}$$

The bisection method is straightforward to implement, and will always converge on the solution, provided that the function is continuous over the whole range between the end points. However, there are a number of important shortcomings that must be considered. First, it is relatively slow compared with other techniques since the convergence proceeds as described in equation (3.5) regardless of the shape of the function. Second, serious problems arise if there is more than one solution within the starting range for the search. For example, if there are two solutions in the range, then there will be two sign changes in the function. As such, the signs of $f(a)$ and $f(b)$ will be identical and the search will fail. Therefore, great care must be taken when applying the bisection method to problems where the possible range of each solution is unknown.

3.2 Newton–Raphson root finding

The Newton–Raphson technique is an alternative method for finding the solutions of $f(x) = 0$. However, in contrast to the bisection method, it does not search for a solution between

two limits. Instead, an initial estimate of the solution x_0 is used as the starting point. The algorithm then uses the gradient of the function at that point to estimate where the function crosses the axis. The algorithm proceeds as follows:

(1). Compute the function $f(x_0)$ and its gradient $f'(x_0)$ at the starting point.

(2). Use the gradient to extrapolate the function back to the x-axis to obtain a new estimate of the solution:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \tag{3.8}$$

(3). Repeat until the solution is found to the desired precision $\delta x = x_n - x_{n-1}$.

The Newton–Raphson method works by approximating the function as a straight line passing through the current point at each iteration. It produces an estimate of the solution directly at each iteration, whereas the bisection method merely reduces the uncertainty at each step. As such, the Newton–Raphson method can find solutions extremely rapidly in many cases.

However, there are again a number of important limitations. First, the algorithm is *unbounded*, meaning that the new estimate of the solution at each step is not restricted to any particular range of values. As such, it is entirely possible that x_n is actually further from the solution than the previous x_{n-1}, leading to overshoot or divergence from the solution. Even worse, the new estimate can lie outside the domain of the function being solved. For example, a negative solution could be predicted for $\ln(x) = 0$ depending on the starting estimate. The unbounded nature of the search algorithm also means that it is not possible to select a particular solution in preference to any other. In fact, the algorithm can converge on *any* solution. As a rule of thumb, though, the Newton–Raphson method works very well for smoothly varying functions with widely separated solutions.

Another issue is that many functions cannot be differentiated analytically. This is a particular problem when calculations are performed on sampled data, as is the case for the Schrödinger equation solutions of complicated structures later in this chapter. In this case, the *secant method* can be used, which approximates the derivative of a function at point x_n using data from the previous iteration:

$$f'(x) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} \tag{3.9}$$

The method then proceeds according to equation (3.8) as before. The only change here is that two initial value estimates are required, rather than one, in order to generate an initial estimate of the derivative.

Finally, it is important to note that there are many other root-finding algorithms beside the bisection and Newton–Raphson approaches. One very attractive option is the Brent algorithm [2], which provides a bounded, yet very rapid search by attempting to use the secant method where possible but falls back on the bisection method if necessary. Although the implementation of the Brent algorithm is beyond the scope of this discussion, it is widely available in free software libraries such as the *GNU Scientific Library* [3].

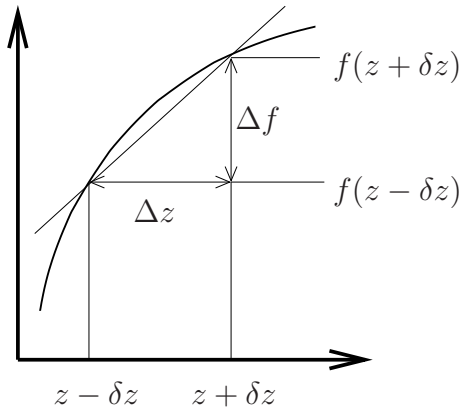


Figure 3.2: Numerical approximation to the first derivative of a function

3.3 Numerical differentiation

General numerical solutions for the Schrödinger equation (Section 3.4) take an arbitrary confining potential $V(z)$ as the starting point and compute the wave functions of the confined quantum states. Since the analytical form of $V(z)$ is generally unknown, it is no longer possible to perform analytical studies of the system. Instead, *samples* of $V(z)$ may be taken at a number of spatial locations with an interval δz between points. It is then possible to compute solutions to the Schrödinger equation using discrete mathematical techniques. One key element in these calculations is the ability to determine the derivative of a function using discrete sampled data.

Figure 3.2 shows that the first derivative of a function $f(z)$ is defined as:

$$\lim_{\Delta z \to 0} \frac{\Delta f}{\Delta z} = \frac{\mathrm{d}f}{\mathrm{d}z} \tag{3.10}$$

It suits the purpose here to retain the approximate form, i.e.

$$\frac{\mathrm{d}f}{\mathrm{d}z} \approx \frac{\Delta f}{\Delta z} = \frac{f(z+\delta z) - f(z-\delta z)}{2\delta z} \tag{3.11}$$

Hence, the second derivative follows as:

$$\frac{\mathrm{d}^2 f}{\mathrm{d}z^2} \approx \frac{\left.\frac{\mathrm{d}f}{\mathrm{d}z}\right|_{z+\delta z} - \left.\frac{\mathrm{d}f}{\mathrm{d}z}\right|_{z-\delta z}}{2\delta z} \tag{3.12}$$

By using the finite-difference forms in equation (3.11) for the first derivatives, then:

$$\frac{\mathrm{d}^2 f}{\mathrm{d}z^2} \approx \frac{\left[\frac{f(z+2\delta z)-f(z)}{2\delta z}\right] - \left[\frac{f(z)-f(z-2\delta z)}{2\delta z}\right]}{2\delta z} \tag{3.13}$$

$$\therefore \frac{\mathrm{d}^2 f}{\mathrm{d}z^2} \approx \frac{f(z+2\delta z) - 2f(z) + f(z-2\delta z)}{(2\delta z)^2} \tag{3.14}$$

As δz is an as yet undefined small step length along the z-axis, and as it only appears in equation (3.14) with the factor 2, then this finite-difference representation of the second derivative can be simplified slightly by substituting δz for $2\delta z$, i.e.

$$\frac{\mathrm{d}^2 f}{\mathrm{d}z^2} \approx \frac{f(z+\delta z) - 2f(z) + f(z-\delta z)}{(\delta z)^2} \tag{3.15}$$

3.4 Discretised Schrödinger equation

The time-independent Schrödinger equation may now be reformulated as a *difference equation*, in which the functions are expressed using discrete samples over a number of spatial points rather than as continuous analytical expressions.

As a starting point, consider a general, but simple (constant-mass) form for the time-independent Schrödinger equation, the analytical solutions of which have been extensively studied in the previous chapter:

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial z^2}\psi(z) + V(z)\psi(z) = E\psi(z) \tag{3.16}$$

where the one-dimensional potential $V(z)$ will remain undefined and again $\psi(z)$ is the wave function representing the particle of interest *while under the effective mass and envelope function approximations*.

Using the finite-difference form for the second derivative in the original Schrödinger equation and taking the step length δz as sufficiently small that the approximation is good, i.e. drop the '$\approx$' in favour of '=', then:

$$-\frac{\hbar^2}{2m^*}\left[\frac{\psi(z+\delta z) - 2\psi(z) + \psi(z-\delta z)}{(\delta z)^2}\right] + V(z)\psi(z) = E\psi(z) \tag{3.17}$$

This can be rearranged so that each sample of $\psi(z)$ appears in the sequence:

$$a\,\psi(z-\delta z) + b(z)\,\psi(z) + c\,\psi(z+\delta z) = E\psi(z), \tag{3.18}$$

where

$$a = c = -\frac{\hbar^2}{2m^*(\delta z)^2} \quad \text{and} \quad b(z) = \frac{\hbar^2}{m^*(\delta z)^2} + V(z) \tag{3.19}$$

This equation describes how the wave function at point z is related to its value at neighbouring points, in terms of the potential, the mass and the energy of the state. It is important to note that this is a completely general equation, which can be applied to *any* one-dimensional confining potential. The following sections describe methods to solve this difference equation to find $\psi(z)$ and E.

3.5 Shooting method

The discretised Schrödinger equation (equation (3.18)) can be written as:

$$\psi(z+\delta z) = \frac{1}{c}[E - b(z)]\psi(z) - \frac{a}{c}\psi(z-\delta z) \tag{3.20}$$

Since $a = c$ (for constant effective mass), this can be simplified to obtain:

$$\psi(z + \delta z) = \frac{1}{c}[E - b(z)]\psi(z) - \psi(z - \delta z) \tag{3.21}$$

Equation (3.21) implies that if the wave function is known at the two points $z - \delta z$ and z, then the value of the wave function at $z + \delta z$ can be calculated for any energy E. This iterative equation forms the basis of a standard method of solving differential equations numerically, and is known as the *shooting method* [4].

Using two known values of the wave function $\psi(z - \delta z)$ and $\psi(z)$, a third value, i.e. $\psi(z + \delta z)$, can be predicted. Using this new point $\psi(z + \delta z)$, together with $\psi(z)$, and by making the transformation $z + \delta z \rightarrow z$, a fourth point, $\psi(z + 2\delta z)$, can be calculated, and so on. Hence the complete wave function can be deduced for any particular energy. The solutions for stationary states have wave functions which satisfy the standard boundary conditions, i.e.

$$\psi(z) \rightarrow 0 \quad \text{and} \quad \frac{\partial}{\partial z}\psi(z) \rightarrow 0, \quad \text{as } z \rightarrow \pm\infty \tag{3.22}$$

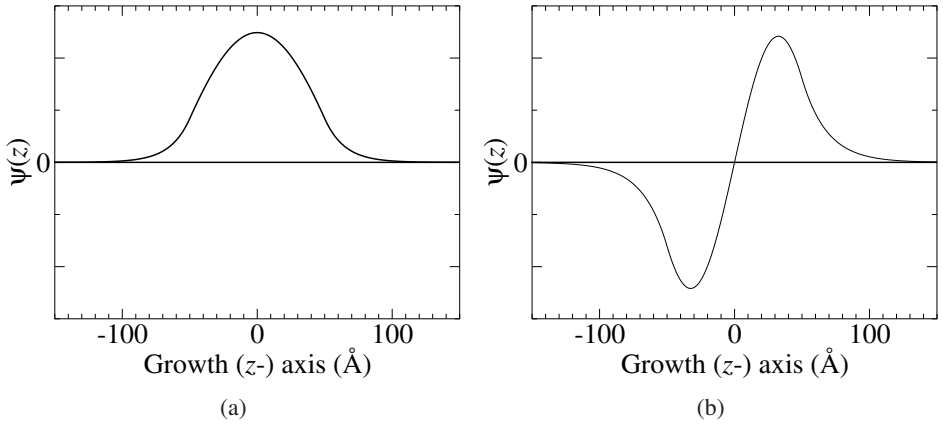


Figure 3.3: Wave functions for the (a) even-parity ground state and (b) odd-parity first excited state in a symmetrical quantum well

The first two values of the wave function necessary to start off the procedure *can* be deduced by using simple symmetry arguments, namely that if the potential $V(z)$ is symmetric, then the eigenstates must be either symmetric (even parity) or anti-symmetric (odd parity). If the state of interest has odd parity, e.g. the first excited state of a symmetric quantum well, as illustrated in Fig. 3.3(b), then the wave function at the centre of the well (call this the origin $z = 0$ for now) must be zero. Correspondingly, a small displacement along the growth (z-)direction must yield a finite value for the wave function. The actual magnitude is not relevant since the energy eigenvalues of the linear Schrödinger equation

are unchanged if the wave function is scaled by any given number. Hence in this case, the following starting conditions could be chosen:

$$\psi(0) = 0, \qquad \psi(\delta z) = 1 \tag{3.23}$$

Given this, it remains to find the eigenvalue energy E. As stated above, the value of E corresponding to a stationary state, or more specifically in this case a confined state within a well potential, is that value which produces a wave function conforming to the standard boundary conditions. As E is an unknown in equation (3.21), then ψ is really a function of both position z *and* energy E, since given the starting conditions a wave function can be generated for any E—although it will not always tend to zero at infinity and be a stationary state. Thus the wave function should be written as $\psi(z, E)$, in which case solutions are sought to the equation:

$$\psi(\infty, E) = 0 \tag{3.24}$$

which can be found by using standard techniques such as the Newton–Raphson iteration (Section 3.2).

If the eigenstate of interest has even parity, e.g. the ground state of a symmetric quantum well, as shown in Fig. 3.3(a), then new starting conditions must be deduced. In particular, as the value of the wave function at the origin is non-zero, then $\psi(0) = 1$ can be chosen. In addition, as the wave function is symmetric then $\psi(-\delta z) = \psi(+\delta z)$. Substituting both of these expressions into equation (3.21) then gives:

$$\psi(+\delta z) = \frac{1}{c}[E - b(0)] \times 1 - \psi(+\delta z) \tag{3.25}$$

$$\therefore \psi(+\delta z) = \frac{m^*(\delta z)^2}{\hbar^2}\,[V(0) - E] + 1 \tag{3.26}$$

3.6 Generalised initial conditions

While such initial (starting) conditions for symmetric potentials are very useful, they are restrictive in that there are many systems that do not have symmetric potentials, perhaps the most obvious being the quantum well with an electric field applied (see Section 2.10). To deduce more general starting conditions for the iterative equation (equation (3.21)), note that at the moment particles are always confined by potential barriers at the edges of the system. Hence, as stated above many times, all wave functions satisfy the standard boundary conditions:

$$\psi(z) \to 0 \quad \text{and} \quad \frac{\partial}{\partial z}\psi(z) \to 0, \quad \text{as } z \to \pm\infty \tag{3.27}$$

In addition, learning from the analytical solutions of the previous chapter, the wave function decays exponentially into the end barriers (see Fig. 3.4). The decay constant κ is also given by:

$$\kappa = \frac{\sqrt{2m^*(V - E)}}{\hbar} \tag{3.28}$$

Hence, upon choosing the first value of the wave function $\psi(z - \delta z) = 1$,

$$\psi(z - \delta z) = \psi(z) \exp\left(-\kappa|\delta z|\right) \tag{3.29}$$

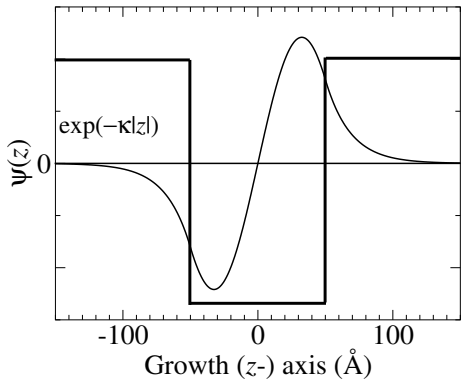


Figure 3.4: Exponential decay of the wave function into the end barrier

Therefore, given the starting value $\psi(z - \delta z)$, the next value $\psi(z)$ can be calculated, thus implementing the boundary condition of exponential growth.

In practice, however, these boundary conditions *can* be unreliable, not because of the starting conditions they impose, but because of the boundary conditions that must be sought at the other end of the potential structure. The mathematics implies that energies must be sought for which ψ tends to zero; however, the original choice of $\psi = 1$ at the start of the potential structure immediately implies an asymmetry. Generally this asymmetry is very small, but not always. In order to correct this asymmetry, the (at first sight bizarre) starting conditions are chosen as:

$$\psi(z - \delta z) = 0 \quad \text{and} \quad \psi(z) = 1 \tag{3.30}$$

Such starting conditions for the shooting method solution of the Schrödinger equation are important because of their generality. They are applicable to all potential profiles, whether symmetric or not, whether the outer barrier is flat or not, and whether the eigenstate of interest is symmetric, anti-symmetric or without definite parity.

These conditions can be partially justified mathematically. As multiplying an eigenstate (wave function ψ) by a constant does not affect the eigenvalue (the energy E), if the first wave function point, $\psi(z - \delta z)$, was taken as a small but finite value (say, $\delta\psi$), then the second starting point, $\psi(z)$, could be given any value larger than $\delta\psi$ (say, $N\delta\psi$), where N is a large number, while still giving exponential growth and without changing the energy eigenvalue. The third value of the wave function simply follows from the shooting equation (equation (3.21)) as follows:

$$\psi(z + \delta z) = \left\{ \frac{2m^*}{\hbar^2}(\delta z)^2 \left[V(z) - E\right] + 2 \right\} N\delta\psi - \delta\psi \tag{3.31}$$

$$\therefore \psi(z + \delta z) = \left(\left\{ \frac{2m^*}{\hbar^2}(\delta z)^2 \left[V(z) - E\right] + 2 \right\} N - 1 \right) \delta\psi \tag{3.32}$$

These new starting conditions of 0 and 1 merely represent the limit of large N.

While the above only offers partial justification, full vindication will be given by the results of the convergence tests in the following sections.

3.7 Practical implementation of the shooting method

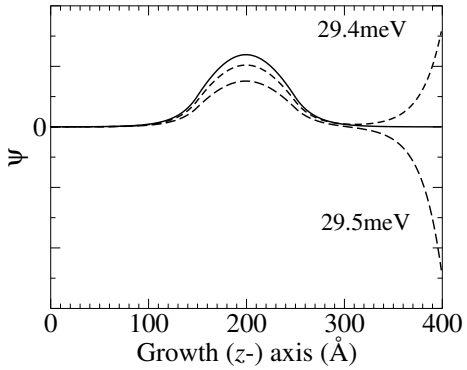


Figure 3.5: Numerically obtained wave functions above and below the true solution at $E = 29.43\,\mathrm{meV}$, for an electron in a 150 Å $Ga_{0.8}Al_{0.2}As$/100 Å GaAs/150 Å $Ga_{0.8}Al_{0.2}As$ single quantum well

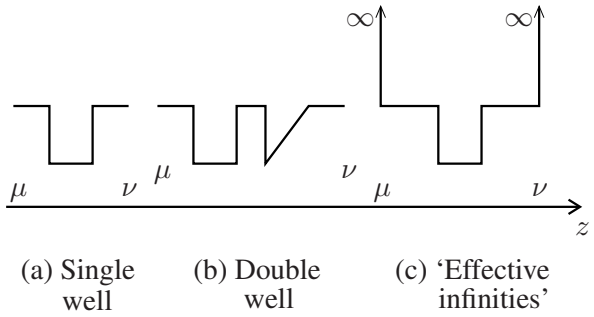


Figure 3.6: Schematic illustration of the extent of the potential $V(z)$, where $\mu < z < \nu$

A two-stage search method may be used to locate the solutions of equation (3.21). First, the range of energies under consideration is divided coarsely into a set of discrete cells. In many cases, the states of interest are those that lie between the minimum and maximum values of the confining potential (i.e. the bound states). The size of each energy cell, δE, is selected to be smaller than the minimum energy spacing between states.

The sign of $\psi(z \to \infty)$ is computed at the maximum and minimum energy of each cell, and a sign change indicates that a solution lies within that range. Figure 3.5 illustrates a true solution at $E = 29.43\,\mathrm{meV}$, which satisfies the boundary condition at $z \to \infty$, and two incorrect solutions at energies slightly above and below the correct value. The wave functions

that correspond to the incorrect solutions are shown to diverge in opposite directions. Having detected the existence of a solution within a given cell, a Newton–Raphson iteration is then used to converge on the true solution to the desired precision.

In practice, the wave function iteration is begun a finite distance to the left of the first well in the potential and is halted a finite distance to the right of the last well (see Fig. 3.6). The potential, which can be merely a list of numbers specifying $V(z)$ at regular intervals along the z-axis, defines the step length, the extent of the wave functions, and the effective infinities. The latter are the points by which the wave function is considered to have converged towards zero at both limits of the potential profile. Seeking solutions based on this criterion is, in effect, equivalent to applying infinite potentials a sufficient distance into the outer barriers of the finite structure, as shown in Fig. 3.6(c). The positions of these effective infinities, μ and ν, which are the lower and upper limits of the z-domain, should be chosen to be of sufficient extent so as not to affect the eigenvalues. Defining this mathematically:

$$\lim_{\mu\to-\infty} E = E_n \quad \text{and} \quad \lim_{\nu\to+\infty} E = E_n \tag{3.33}$$

where E_n are the set of true eigenenergies.

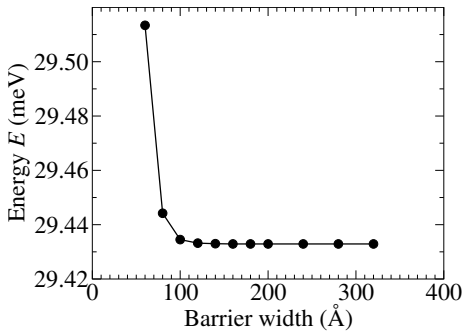


Figure 3.7: Effect of barrier width (effective infinities, μ and ν) on energy

In Fig. 3.7 the left- and right-hand barriers are varied in thickness to illustrate the effect on the solution for the energy produced by this numerical method, for the single quantum well as discussed above. Clearly, the energy converges to a constant, which is the true stationary state—in this case, the ground state E_1. Thus if the energy is the only motivation for the calculation, any barrier width beyond, say, $150\,\text{Å}$ will suffice. Note, however, that this can vary depending on the barrier height, well width and carrier effective mass.

If the wave function is also desired, e.g. to be used as an input for calculating another property, such as the exciton binding energy or an electron–phonon scattering rate, then greater care has to be taken. While very large outer barrier layers can be chosen to ensure convergence of the eigenenergy, this can have a detrimental effect on the wave function, as can be seen in Fig. 3.8. At too narrow a barrier width, e.g. the $60\,\text{Å}$ case, although the energy may be returned to within $1\,\text{meV}$, the wave function does not satisfy the second of the standard boundary conditions, i.e. $\psi' = 0$. With a $200\,\text{Å}$ barrier, which corresponds to

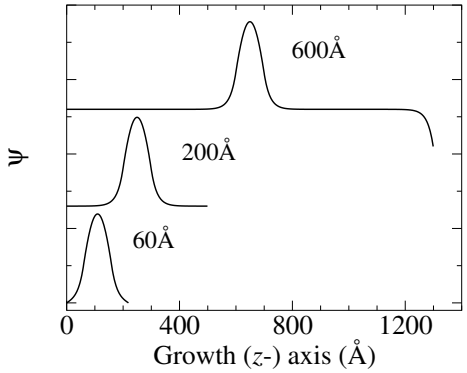


Figure 3.8: Effect of barrier width (effective infinities, μ and ν) on the solution to the wave function

just after the energy has converged, the wave function is as expected and satisfies both the standard boundary conditions—this represents the optimal barrier width. However, at very large barrier widths, such as 600 Å as illustrated, the wave function is beginning to diverge, and, as stated above, although the energy obtained from such a solution is reliable, the wave function cannot be used to calculate other properties.

The energy E can always be obtained to far higher accuracy than would be required for comparison of, say, electron confinement energies with photoluminescence data. Computer codes typically assign 16-figure accuracy to a double-precision number, and the shooting method can be used to determine E to 12 of these, for example. However, even at this high accuracy the wave function at the end of an iteration, i.e. the value of $\psi(\infty, E) = \psi(\nu, E)$, cannot be made equal to zero for large barrier widths, simply because E cannot be stored to enough significant figures. If 32-figure accuracy could be used in the computing, then the wave function could be returned realistically for larger barrier widths.

The wave functions obtained from this numerical method are not normalised, i.e. they do not satisfy:

$$\int_{\text{all space}} \psi^*(z)\psi(z)\,\mathrm{d}z = 1 \tag{3.34}$$

This can easily be achieved with the following transformation:

$$\psi(z) \rightarrow \frac{\psi(z)}{\sqrt{\int_{\text{all space}} \psi^*(z)\psi(z)\,\mathrm{d}z}} \tag{3.35}$$

This numerical solution to a still simplistic Schrödinger equation is of use as it enables comparison with a number of curved potentials that have analytical solutions, such as the parabolic and the Pöschl–Teller potentials.

3.8 Heterojunction boundary conditions

In the previous chapter it was shown that the conditions for matching solutions at interfaces between dissimilar materials were fixed as soon as the Hamiltonian was decided. This is also reflected here within the shooting method. Consider again the shooting equation for the constant-mass Hamiltonian ($\frac{1}{m^*}\mathcal{P}_z\mathcal{P}_z$) (equation (3.21)):

$$\psi(z+\delta z) = \left\{ \frac{2m^*}{\hbar^2}(\delta z)^2 \left[V(z) - E\right] + 2 \right\} \psi(z) - \psi(z-\delta z) \tag{3.36}$$

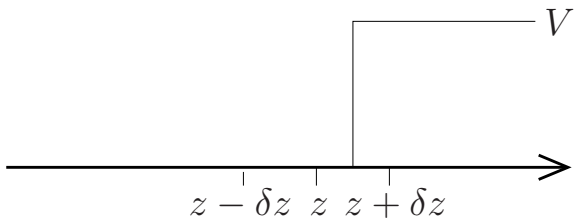


Figure 3.9: Calculating a new wave function point across a discontinuity in the potential

Consider also that the two known points of the wave function, $\psi(z-\delta z)$ and $\psi(z)$, lie within a quantum well with $V(z) = 0$, but the new point to be calculated lies across a heterojunction in a barrier (as in Fig. 3.9). Then:

$$\psi(z+\delta z) = \left[\frac{2m^*}{\hbar^2}(\delta z)^2 \left(-E\right) + 2 \right] \psi(z) - \psi(z-\delta z) \tag{3.37}$$

Then in the limit of decreasing step length δz, i.e. eliminating terms in $(\delta z)^2$:

$$\psi(z+\delta z) = 2\psi(z) - \psi(z-\delta z) \tag{3.38}$$

$$\therefore \psi(z+\delta z) - \psi(z) = \psi(z) - \psi(z-\delta z) \tag{3.39}$$

which implies continuity of the *derivative* across a heterojunction, as expected for this constant effective mass Hamiltonian.

3.9 Matrix solutions of the discretised Schrödinger equation

Section 3.7 illustrated an important shortcoming of the shooting method: that the accuracy of solutions is highly dependent on the boundary conditions selected. This is a fundamental issue that arises from the limited precision with which numbers are stored in computer memory.

Additionally, some thought must be given to the speed of the shooting method. If the structure contains N_z spatial points (typically with a step length of $\sim$1 Å), then N_z calculations are required to compute the complete wave function at a given energy. The

calculation must then be completed at least once for each energy cell in the system. The energy spacing between states (and hence the number of cells to be inspected) depends upon the system under consideration, although clearly longer structures exhibit smaller energy spacings. As such, the shooting method can become quite slow when large structures with many states are being considered.

These shortcomings may be overcome by using linear algebra techniques to solve equation (3.18) directly rather than by searching a set of energy cells. First, rewrite the equation using more compact notation:

$$a_i\,\psi_{i-1} + b_i\,\psi_i + c_i\,\psi_{i+1} = E\psi_i \tag{3.40}$$

where the subscript denotes the *index* of each sample of the wave function, rather than explicitly writing out its location, such that $\psi_0 = \psi(0)$, $\psi_1 = \psi(\delta z)$ and so on. The coefficients are then given by:

$$a_{i+1} = c_i = -\frac{\hbar^2}{2m^*(\delta z)^2} \quad \text{and} \quad b_i = \frac{\hbar^2}{m^*(\delta z)^2} + V_i \tag{3.41}$$

Now, taking the standard boundary conditions, $\psi_0 = \psi_{N+1} = 0$, at the points just outside the heterostructure, equation (3.40) can be expressed at each sample point using a system of linear equations:

$$\begin{aligned} b_1\psi_1 + c_1\psi_2 &= E\psi_1 \\ a_2\psi_1 + b_2\psi_2 + c_2\psi_3 &= E\psi_2 \\ &\cdots \\ a_{N-1}\psi_{N-2} + b_{N-1}\psi_{N-1} + c_{N-1}\psi_N &= E\psi_{N-1} \\ a_N\psi_{N-1} + b_N\psi_N &= E\psi_N \end{aligned} \tag{3.42}$$

This system of equations can be represented more succinctly in matrix form:

$$\mathsf{H}\boldsymbol{\psi} = E\boldsymbol{\psi} \tag{3.43}$$

where H is a matrix containing all the coefficients:

$$\mathsf{H} = \begin{pmatrix} b_1 & c_1 & 0 & \cdots & 0 \\ a_2 & b_2 & c_2 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \cdots & a_{N-1} & b_{N-1} & c_{N-1} \\ 0 & \cdots & 0 & a_N & b_N \end{pmatrix} \tag{3.44}$$

and $\boldsymbol{\psi}$ is a column vector containing all the samples of the wave function in their correct order:

$$\boldsymbol{\psi}^{\mathrm{T}} = \{\psi_1, \psi_2, \ldots, \psi_N\} \tag{3.45}$$

Equation (3.43), therefore, represents a matrix eigenvalue problem, which can be solved directly to locate *all* the energy eigenvalues of the system simultaneously, along with the

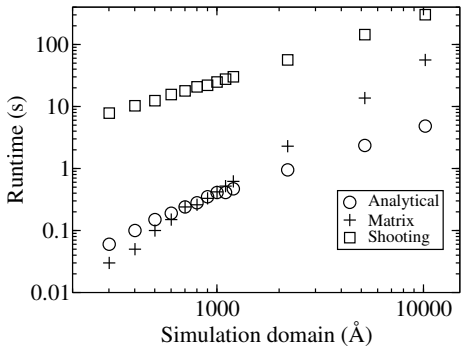


Figure 3.10: Time taken to solve the Schrödinger equation for a finite square GaAs well surrounded by a pair of 100 Å AlAs barriers, assuming parabolic dispersion. The domain length is defined as the sum of the barrier and well widths, and a 1 Å separation between sampling points was used. Results are shown for the analytical solver of Section 2.7, the shooting method of Section 3.5 and the matrix method of Section 3.9. All calculations were undertaken on a desktop computer with a 3.2 GHz AMD Phenom II X4-955 processor and 16 GB RAM

corresponding wave functions. It is also important to note that the matrix H only contains non-zero terms along its three main diagonals, and is also symmetrical since $a_{i+1} = c_i$. This type of *real symmetric tridiagonal* problem can be solved using a range of extremely fast numerical algorithms. Although a detailed description of the algorithms is beyond the scope of this book, they are widely available in free software libraries including LAPACK [5] and the *GNU Scientific Library* [3], or can be implemented quite simply. Figure 3.10 illustrates the speed of the matrix method, and confirms that it is considerably faster than the shooting method for a square well. Indeed, for simulation domains smaller than 700 Å, the matrix method is faster than the analytical solution!

The matrix methods also locate solutions to machine precision (i.e. the highest precision that can be represented by a computer) directly rather than through an iterative search process, meaning that numerical errors are less significant than in the shooting method. Furthermore, since the boundary conditions are implicitly included within the H matrix itself, the computed wave functions cannot diverge at the boundaries when this approach is used.

The accuracy of the numerical method can be tested by inspecting whether fundamental mathematical properties of the wave functions are satisfied. Recalling that since the Schrödinger equation is an eigenvalue problem, all wave functions (the eigenvectors) must be *orthogonal* to each other. Furthermore, the normalisation condition (equation (2.14)) must be satisfied. These properties are together known as the *orthonormality condition* of wave functions, and can be tested by solving the *overlap integral*:

$$O = \langle \psi_n | \psi_{n'} \rangle = \int_{-\infty}^{+\infty} \psi_n^* \psi_{n'} \, \mathrm{d}z \tag{3.46}$$

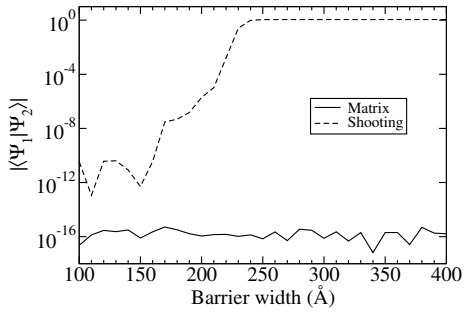


Figure 3.11: Magnitude of overlap integral for the first two states in a 150 Å GaAs quantum well as a function of the barrier width, assuming parabolic dispersion. A 1 Å separation between sampling points was used. Results are shown for the shooting method of Section 3.5 and the matrix method of Section 3.9

If ψ_n and $\psi_{n'}$ are eigenstates of the same Hamiltonian, whether they be wave or envelope functions, then the orthonormality condition states that:

$$\langle\psi_n|\psi_{n'}\rangle = \delta_{nn'} \tag{3.47}$$

Figure 3.11 shows that the overlap between the first pair of states in a square well is effectively zero when calculated using the matrix method, irrespective of barrier width. However, as expected from Section 3.7, the shooting method gives good results up to barrier widths of ~ 150 Å but rapidly becomes inaccurate as the barrier width increases.

These investigations show that matrix methods for solving the discretised Schrödinger equation are typically much faster, more robust and more accurate than shooting method approaches and are generally a better choice for longer structures.

3.10 The parabolic potential well

The parabolic potential well is a good testing ground for the accuracy of the numerical method introduced, in that it allows progression from the standard step potentials that have been dealt with over and over again, to a curved profile, but still with analytical solutions to compare with.

The parabolic potential well is a direct analogy to the harmonic oscillator of classical mechanics in that the potential is, of course, proportional to the square of the displacement, i.e., following Eisberg [6]:

$$V(z) = C\frac{z^2}{2} \tag{3.48}$$

where C is a constant. If a particle of mass m is displaced from the equilibrium position by an amount Δz, then the restoring force is proportional to $-\partial V/\partial z = -Cz$, thus causing oscillations of amplitude Δz and angular frequency:

$$\omega = \sqrt{\frac{C}{m}} \tag{3.49}$$

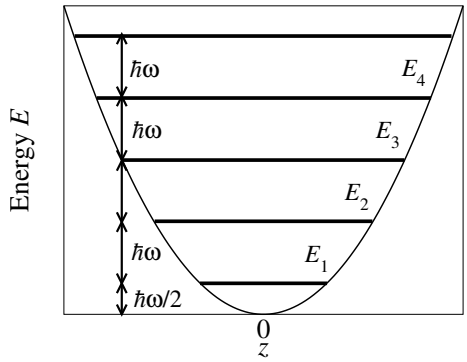


Figure 3.12: The quantum mechanical harmonic oscillator

The behaviour of a quantum particle, such as an electron, within such a potential is quite different. The Schrödinger equation for the constant-mass case would then become:

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial z^2}\psi + C\frac{z^2}{2}\psi = E\psi \tag{3.50}$$

Substituting ω for C, then:

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial z^2}\psi + m^*\omega^2\frac{z^2}{2}\psi = E\psi \tag{3.51}$$

Table 3.2 Convergence test of the shooting method solution for a parabolic quantum well; the energy eigenvalues E_n for the lowest 10 confined states—note that the value for E_1 implies $\hbar\omega = 871.876\,\mathrm{meV}$

State index, n	Eigenvalue, E_n (meV)	$E_n/(2E_1)$
1	435.938	0.500
2	1307.809	1.500
3	2179.671	2.500
4	3051.522	3.500
5	3923.341	4.500
6	4794.990	5.500
7	5665.755	6.498
8	6532.518	7.492
9	7382.747	8.468
10	8156.309	9.355

The method of solution of this eigenvalue problem has been covered many times in standard texts, for example [6, 7] or alternatively [8]. It is sufficient here to quote the result,

which says that the energy levels are quantised and given by:

$$E_{n+1} = \left(n + \frac{1}{2}\right)\hbar\omega \tag{3.52}$$

where the subscript $(n+1)$ has been introduced in order for the notation to agree with that developed so far, i.e. the ground state being labelled '1'.

Figure 3.12 illustrates the solution; the ground state has an energy $\hbar\omega/2$ even at the absolute zero of temperature, thus coining the term the *zero-point energy*. Above this the equal-spaced energy steps form a ladder, which has prompted suggestions for exploitation in nonlinear optics [9].

Focusing on parabolic quantum wells within semiconductor multilayers, it has to be acknowledged that, as only finite potentials are available through the standard band offsets, the potential profile will resemble that shown in Fig. 3.13, i.e. a *finite parabolic quantum well*, rather than that without limit as shown in Fig. 3.12.

Taking the central well to be of width a and the outer barriers to have widths b, the parabolic quantum well can then be specified by the variation in the alloy component x along the z-axis (in $Ga_{1-x}Al_xAs$, say). In particular, if x is allowed to vary from a minimum value x_{min} to a maximum x_{max}, then the well profile is given by the following quadratic relationship:

$$x(z) = x_{\text{min}} + \frac{\left[z - (b + a/2)\right]^2 (x_{\text{max}} - x_{\text{min}})}{(a/2)^2} \tag{3.53}$$

Table 3.2 displays the energies of the lowest ten levels within a $Ga_{1-x}Al_xAs$ parabolic quantum well. The layer thicknesses have been taken as $a = b = 100\,\text{Å}$, and the effective mass has been taken as a constant, $0.067m_0$; however, the maximum alloy concentration, x_{max}, has been taken artificially as high as 10, in order to produce a large number of confined levels. As $E_1 = \hbar\omega/2$, the third column of data displays the 'half-integer' $n + \frac{1}{2}$. Clearly, the agreement with the analytical theory is exact for the lower states, although as the levels approach the top of the barrier, they experience the finiteness of the potential, and hence some discrepancy with the infinite parabola solution arises.

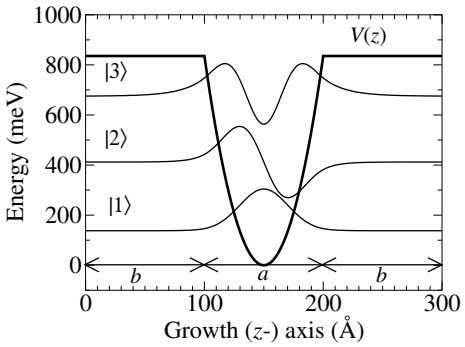


Figure 3.13: The three confined eigenstates of a finite parabolic quantum well

The energy step $\hbar\omega$ between the levels is a useful quantity for describing the properties of the well, even though ω doesn't represent the angular frequency of oscillation. With this in mind, consider the conversion of the alloy concentration of equation (3.53) into a potential profile (see Appendix A). The conduction-band profile of $Ga_{1-x}Al_xAs$ would then become:

$$V(z) = \Delta V_{\text{CB}}\; x(z)\left(E_{\text{g}}^{\text{AlAs}} - E_{\text{g}}^{\text{GaAs}}\right) \tag{3.54}$$

Recalling that $\partial V/\partial z = Cz$, then:

$$C = \frac{1}{z}\frac{\partial V}{\partial z} = \frac{1}{z}\Delta V_{\text{CB}}\;\left(E_{\text{g}}^{\text{AlAs}} - E_{\text{g}}^{\text{GaAs}}\right)\frac{\partial}{\partial z}x \tag{3.55}$$

Ignoring the origin shift, $b - a/2$:

$$C = \frac{1}{z}\Delta V_{\text{CB}}\;\left(E_{\text{g}}^{\text{AlAs}} - E_{\text{g}}^{\text{GaAs}}\right)\;\;\frac{8z\left(x_{\text{max}} - x_{\text{min}}\right)}{a^2} \tag{3.56}$$

$$\therefore C = \Delta V_{\text{CB}}\;\left(E_{\text{g}}^{\text{AlAs}} - E_{\text{g}}^{\text{GaAs}}\right)\;\;\frac{8\left(x_{\text{max}} - x_{\text{min}}\right)}{a^2} \tag{3.57}$$

Then ω follows from equation (3.49) as:

$$\omega = \sqrt{8\Delta V_{\text{CB}}\left(E_{\text{g}}^{\text{AlAs}} - E_{\text{g}}^{\text{GaAs}}\right)\left(x_{\text{max}} - x_{\text{min}}\right)}\;\;\frac{1}{a\sqrt{m^*}} \tag{3.58}$$

By using the values corresponding to the data in Table 3.2 ($V_{\text{CB}} = 0.67$, $E_{\text{g}}^{\text{AlAs}} - E_{\text{g}}^{\text{GaAs}} = 1247\,\text{meV}$, $x_{\text{max}} = 10$, $x_{\text{min}} = 0$, $a = 100$ Å and $m^* = 0.067m_0$), $\hbar\omega = 871.879\,\text{meV}$, which is in excellent agreement with the numerical result as displayed in the table.

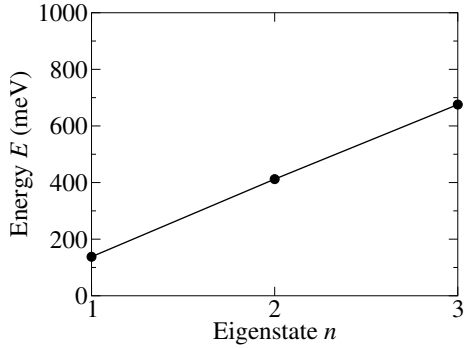


Figure 3.14: Eigenenergies for a GaAs/AlAs parabolic quantum well

Figures 3.13 and 3.14 display the wave functions and energies, respectively, of the three confined states of the more realistic parabolic quantum well, with $x_{\text{max}} = 1$. The wave functions resemble those of the single quantum well, and the energy levels are to all intents and purposes equally spaced.

In summary, it has been demonstrated that the simple shooting method solution of the Schrödinger equation, taken together with the stated starting conditions, can produce energy levels in exact agreement with the analytical values for the curved potential of the parabolic quantum well. In addition, equation (3.58) summarises the behaviour of parabolic wells within semiconductor multilayers. The energy level spacing is inversely proportional to both the well width and the square root of the mass.

3.11 The Pöschl–Teller potential hole

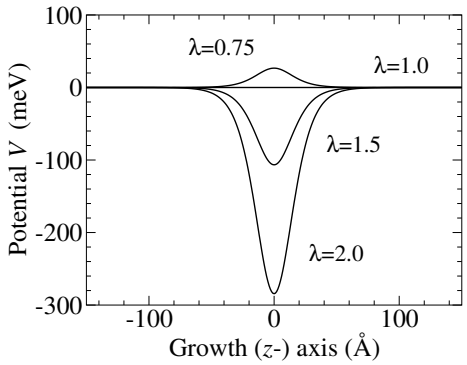


Figure 3.15: Pöschl–Teller potentials as a function of the depth parameter λ for a fixed width parameter α of $0.05\,\text{Å}^{-1}$

The modified Pöschl–Teller potential hole [10], examples of which are shown in Fig. 3.15, is important in that it resembles the profile of a diffused quantum well, but has the advantage of analytical solutions. Therefore as is currently the aim, it will serve well for validation of numerical Schrödinger equation solutions *and* the choice of boundary conditions.

The potential is given by the following [10]:

$$V(z) = -\frac{\hbar^2}{2m^*}\alpha^2\frac{\lambda(\lambda-1)}{\cosh^2\alpha z} \tag{3.59}$$

where α is known as the width parameter and λ as the depth parameter (see Fig. 3.15). Clearly, values of $0 < \lambda < 1$ give potential barriers, $\lambda = 1$ gives a flat band, and potential wells, the area of interest here, are given by values of $\lambda > 1$.

The eigenvalues[1] of the resulting Schrödinger equation are given by [10]:

$$E_n = -\frac{\hbar^2\alpha^2}{2m^*}(\lambda - 1 - n)^2, \qquad n = 0, 1, 2, \ldots \tag{3.60}$$

Table 3.3 compares the analytical solutions given by equation (3.60) for the two lowest-

[1]The eigenvectors also have an analytical form [10], which is available in the QWWAD simulation tools.

Table 3.3 Comparison of the numerical solution with the analytical solution of the Pöschl–Teller potential with a step length δz of 1 Å and the effective mass m^* taken as the GaAs electron mass of $0.067m_0$

	Analytical solution		Numerical solution	
λ	E_1 (meV)	E_2 (meV)	E_1 (meV)	E_2 (meV)
1.5	−35.54088	—	−35.44741	—
2.0	−142.16350	—	−142.17701	—
5.0	−2274.61604	−1279.47152	−2274.92266	−1280.50809
10.0	−11515.24368	−9098.46414	−11516.92404	−9105.64111

energy confined states of the Pöschl–Teller potential holes of Fig. 3.15, with that obtained by the numerical matrix method. The agreement is very good for all values of λ and for both the ground state, of energy E_1, and the first excited state, of energy E_2. However, for the largest value of λ used, which does represent a very deep well, of a magnitude not found in semiconductor heterostructures, the discrepancy between the solutions has increased to 2 and 7 meV for the ground and first excited states, respectively. This discrepancy can be removed by increasing the accuracy of the numerical solution.

3.12 Convergence tests

This is an appropriate point to perform the final convergence test, i.e. the final test of accuracy of the numerical solution of Schrödinger's equation derived in Section 3.9. As discovered in the previous section, while the agreement between the analytical solution of the Pöschl–Teller potential and the numerical solution is generally very good, at times, for certain parameter values, discrepancies can occur. This is often a reflection of the choice of the computational values, for example, and probably most importantly, the resolution in the definition of the potential profile, i.e. the step length δz. The user has complete control over this value, and hence an improved comparison with analytical solutions (should they exist) can be achieved, but, as always, at the inconvenience of increased computation time.

Table 3.4 Comparison of the numerical solution with the analytical solution of the Pöschl–Teller potential as a function of the number (N) of points per angstrom

N (Å^{-1})	Analytical solution		Numerical solution	
	E_1 (meV)	E_2 (meV)	E_1 (meV)	E_2 (meV)
1	−2274.633827	−1279.481528	−2274.940451	−1280.518108
2	−2274.633827	−1279.481528	−2274.710431	−1279.740343
5	−2274.633827	−1279.481528	−2274.646081	−1279.522923
10	−2274.633827	−1279.481528	−2274.636891	−1279.491876
20	−2274.633827	−1279.481528	−2274.634593	−1279.484115
50	−2274.633827	−1279.481528	−2274.633950	−1279.481942
100	−2274.633827	−1279.481528	−2274.633858	−1279.481631

Table 3.4 compares the analytical solutions corresponding to the $\lambda = 5$ potential hole of Table 3.3 with the corresponding numerical solution as a function of the number of potential points $V(z)$ per angstrom along the growth axis. The step length δz in Å is just $1/N$. It can be seen that as the number of points increases the discrepancy between the data decreases (obviously the analytical solutions are not a function of N). At the upper limit of the data in the table, the solutions agree to 7 or 8 significant figures. This is well beyond the accuracy of spectroscopic data, which, in the highest quality of semiconductor heterostructure samples, might yield transition energies to $0.1\,\mathrm{meV}$ (see, for example, [11]), a precision which is reached for $N = 1$.

In summary, for the majority of applications, a step length δz of 1 Å is adequate.

3.13 Extension to variable effective mass

Hitherto, the numerical solution has focused on the constant-mass Schrödinger equation, and this has served the purposes of development and application well. However, in the real world, just as in the analytical solutions before, the Schrödinger equation of real interest accounts for the possibility of the effective mass varying between dissimilar semiconductor layers.

With the aim of generalising the numerical solution for this situation, consider the Schrödinger equation with a spatially dependent mass, i.e.

$$-\frac{\hbar^2}{2}\frac{\partial}{\partial z}\frac{1}{m^*(z)}\frac{\partial}{\partial z}\psi(z) + V(z)\psi(z) = E\psi(z) \tag{3.61}$$

which can be written:

$$\frac{\partial}{\partial z}\frac{1}{m^*(z)}\frac{\partial}{\partial z}\psi(z) = \frac{2}{\hbar^2}\left[V(z) - E\right]\psi(z) \tag{3.62}$$

The variable-mass kinetic energy operator can be expanded directly to give:

$$\frac{\partial}{\partial z}\left[\frac{1}{m^*(z)}\right]\frac{\partial}{\partial z}\psi(z) + \frac{1}{m^*(z)}\frac{\partial^2}{\partial z^2}\psi(z) = \frac{2}{\hbar^2}\left[V(z) - E\right]\psi(z) \tag{3.63}$$

and then:

$$-\frac{1}{[m^*(z)]^2}\frac{\partial}{\partial z}m^*(z)\frac{\partial}{\partial z}\psi(z) + \frac{1}{m^*(z)}\frac{\partial^2}{\partial z^2}\psi(z) = \frac{2}{\hbar^2}\left[V(z) - E\right]\psi(z) \tag{3.64}$$

However, shooting equations derived from this point by expanding the derivatives in terms of finite differences have led to significant computational inaccuracies in systems with a large discontinuous change in the effective mass $m^*(z)$, as occurs, for example, in GaAs/AlAs quantum wells. The source of the inaccuracy is thought to arise from the δ-function nature of $\partial m^*(z)/\partial z$.

A more robust scheme can be derived by expanding the left-hand derivative, $\partial/\partial z$, in equation (3.62), first giving the following:

$$\frac{\frac{1}{m^*(z+\delta z)}\left.\frac{\partial\psi(z)}{\partial z}\right|_{z+\delta z} - \frac{1}{m^*(z-\delta z)}\left.\frac{\partial\psi(z)}{\partial z}\right|_{z-\delta z}}{2\delta z} = \frac{2}{\hbar^2}\left[V(z) - E\right]\psi(z) \tag{3.65}$$

$$\therefore \frac{1}{m^*(z+\delta z)} \left.\frac{\partial \psi}{\partial z}\right|_{z+\delta z} - \frac{1}{m^*(z-\delta z)} \left.\frac{\partial \psi}{\partial z}\right|_{z-\delta z} = \frac{2(2\delta z)}{\hbar^2} [V(z) - E]\, \psi(z) \tag{3.66}$$

Recalling the centred finite-difference expansion for the first derivative, i.e.

$$\left.\frac{\partial f}{\partial z}\right|_z = \frac{f(z+\delta z) - f(z-\delta z)}{2\delta z} \tag{3.67}$$

then:

Table 3.5 Comparison of the numerical solution with the analytical solution for a single quantum well with differing effective masses in the well and barrier

	Analytical solution		Numerical solution	
Well width (Å)	E_1 (meV)	E_2 (meV)	E_1 (meV)	E_2 (meV)
20	126.227914	—	126.204335	—
40	80.111376	—	80.087722	—
60	53.276432	166.522007	53.260451	167.766634
80	37.619825	137.330295	37.609351	137.308742
100	27.884814	106.557890	27.877769	106.535858
120	21.463972	83.606781	21.459068	83.589149
160	13.820474	54.647134	13.817861	54.636614
200	9.629394	38.282383	9.627854	38.275914

$$\frac{1}{m^*(z+\delta z)} \left[\frac{\psi(z+2\delta z) - \psi(z)}{2\delta z}\right] - \frac{1}{m^*(z-\delta z)} \left[\frac{\psi(z) - \psi(z-2\delta z)}{2\delta z}\right] = \frac{2(2\delta z)}{\hbar^2} [V(z) - E]\, \psi(z) \tag{3.68}$$

$$\therefore \left[\frac{\psi(z+2\delta z) - \psi(z)}{m^*(z+\delta z)}\right] - \left[\frac{\psi(z) - \psi(z-2\delta z)}{m^*(z-\delta z)}\right] = \frac{2(2\delta z)^2}{\hbar^2} [V(z) - E]\, \psi(z) \tag{3.69}$$

Gathering terms in $\psi(z)$ on the right-hand side, then:

$$\frac{\psi(z+2\delta z)}{m^*(z+\delta z)} + \frac{\psi(z-2\delta z)}{m^*(z-\delta z)} = \left\{\frac{2(2\delta z)^2}{\hbar^2} [V(z) - E] + \frac{1}{m^*(z+\delta z)} + \frac{1}{m^*(z-\delta z)}\right\} \psi(z) \tag{3.70}$$

Making the transformation $2\delta z \rightarrow \delta z$ then gives:

$$\frac{\psi(z+\delta z)}{m^*(z+\delta z/2)} = \left\{\frac{2(\delta z)^2}{\hbar^2} [V(z) - E] + \frac{1}{m^*(z+\delta z/2)} + \frac{1}{m^*(z-\delta z/2)}\right\} \psi(z) - \frac{\psi(z-\delta z)}{m^*(z-\delta z/2)} \tag{3.71}$$

which is the variable effective mass shooting equation, and is solved according to the boundary conditions (as in Section 3.5). The effective mass m^* can be found at the intermediate points, $z \pm \delta z/2$, by taking the mean of the two neighbouring points at z and $z \pm \delta z$. Clearly, equation (3.72) collapses back to the original form in equation (3.21) when m^* is constant.

Adapting the matrix method (Section 3.9) to handle a spatially varying effective mass only requires the coefficients in the H matrix in equation (3.44) to be modified to use the form taken from equation (3.72):

$$a_{i+1} = c_i = -\frac{\hbar^2}{2m^*_{i+\frac{1}{2}}(\delta z)^2} \quad \text{and} \quad b_i = \frac{\hbar^2}{2(\delta z)^2}\left(\frac{1}{m^*_{i+\frac{1}{2}}} + \frac{1}{m^*_{i-\frac{1}{2}}}\right) + V_i \qquad (3.72)$$

Table 3.6 Comparison of the numerical solution with the analytical solution for a single GaAs quantum well surrounded by $Ga_{0.25}Al_{0.75}As$ barriers, with differing effective masses in the well and barrier

	Analytical solution		Numerical solution	
Well width (Å)	E_1 (meV)	E_2 (meV)	E_1 (meV)	E_2 (meV)
20	270.764969	—	270.357809	—
40	129.450671	512.393113	129.247261	512.047324
60	75.774672	307.023049	75.669905	306.736857
80	49.774624	201.176843	49.715086	200.986873
100	35.204788	141.851288	35.168092	141.724429
120	26.217761	105.411306	26.193670	105.323933
160	16.158835	64.812071	16.146909	64.766509
200	10.949827	43.869233	10.943102	43.842871

Table 3.5 compares the ground state and first excited state energy levels, E_1 and E_2, respectively, calculated with this extended matrix equation, with the analytical solution from Section 2.7.2, for a GaAs quantum well surrounded by $Ga_{0.8}Al_{0.2}As$ barriers. In this series of calculations the step length δz was taken as 1 Å, and it can be seen from the data in the table that the agreement is very good for both the ground state energy E_1 and the first excited state energy E_2 across the range of well widths. The discrepancy between the solutions of the two methods is largest for the excited state of the narrower wells, at which point it is of the order of $1\,\mathrm{meV}$. For the wider wells, the discrepancy reduces to less than $0.1\,\mathrm{meV}$. Such accuracies are entirely acceptable when modelling, for example, experimental spectroscopic data.

As this numerical method for a variable effective mass will be used widely, it is worth performing a few more convergence tests to increase confidence in its applicability. In particular, Table 3.6 repeats the calculations of the previous table but with a much higher barrier Al concentration. The effect of this is twofold: there is an increased difference in potential between the well and barrier, but, more importantly for the present section, there is an increased difference in effective mass. The discrepancies between the analytical solution and the numerical solution are of a similar order as before, for this step length δz of 1 Å.

Table 3.7 Comparison of the numerical solution with the analytical solution for a 20 Å single GaAs quantum well surrounded by $Ga_{0.25}Al_{0.75}As$ barriers, as a function of the number (N) of points per angstrom in the mesh—note the step length $\delta z = 1/N$ (Å)

Number of points, N (Å^{-1})	Analytical solution E_1 (meV)	Numerical solution E_1 (meV)
2	270.764969	270.663106
4	270.764969	270.739499
6	270.764969	270.753649
8	270.764969	270.758601
10	270.764969	270.760894
12	270.764969	270.762139

However, as before, the agreement between the numerical solution and that obtained from the analytical form can be improved by increasing the computational accuracy of the numerical method, i.e. increasing the number of points per angstrom (decreasing δz). This is highlighted in Table 3.7, which compares the solutions from the two methods for a decreasing step length δz.

When using the form for the variable effective mass discretised Schrödinger equation in equation (3.72), the BenDaniel–Duke boundary conditions are 'hard-wired' in. Thus it is unnecessary to repeat the analysis described in Section 3.8 in order to recover them.

3.14 The double quantum well

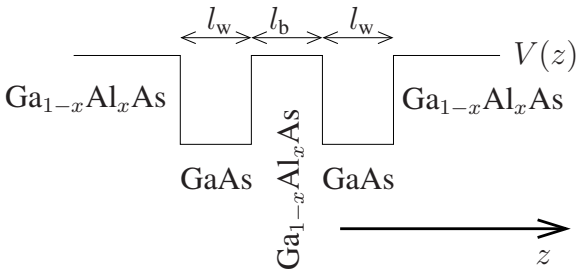


Figure 3.16: The band profile of symmetric GaAs/$Ga_{1-x}Al_xAs$ double quantum well

The convergence tests of the numerical solution to Schrödinger's equation have provided the confidence to apply it to systems for which the analytical solutions have not (in this book!) been developed or do not exist.

The simplest example of the former would be the symmetric double quantum well of Fig. 3.16. The potential function $V(z)$ required for the numerical solution is simply the (in this case) conduction-band edge as given in the figure.

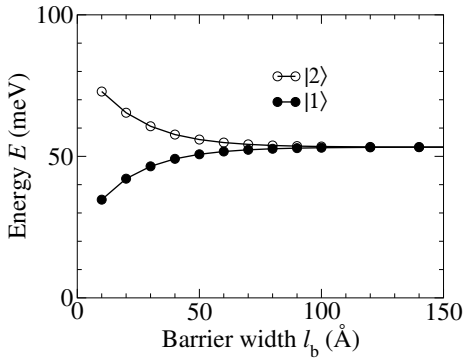


Figure 3.17: The confinement energies of the lowest two states of a symmetric double quantum well as a function of the central barrier width

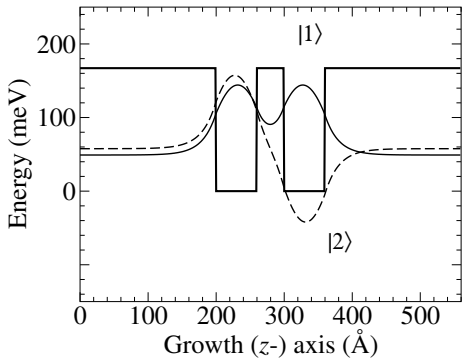


Figure 3.18: The wave functions of the lowest two energy levels of the symmetric double quantum well with a central barrier width of 40 Å

Figure 3.17 displays the results of calculations of the lowest two energy states as a function of the central barrier width for a double quantum well with the Al barrier concentration $x = 0.2$ and a fixed well width $l_{\rm w} = 60$ Å. When the wells are separated by a large distance, the interaction between the eigenstates localised within each well is very small and the wells behave as two independent single quantum wells. However, as illustrated in Fig. 3.17, as the central barrier is decreased, the energy levels interact, with one being forced to higher energies and the other to lower energies. This is directly analogous to the formation of a pair of bonding and anti-bonding orbitals when two hydrogen atoms are brought together to form a hydrogen molecule. In that case, as here, the wave function of the lower (bonding) state is a sum of the wave functions of the separate atoms (wells), and the higher (anti-bonding) state is a difference in wave functions, i.e.

$$\Psi_{\text{bonding}} = \frac{1}{\sqrt{2}}\left(\psi_1 + \psi_2\right), \qquad \Psi_{\text{anti-bonding}} = \frac{1}{\sqrt{2}}\left(\psi_1 - \psi_2\right) \tag{3.73}$$

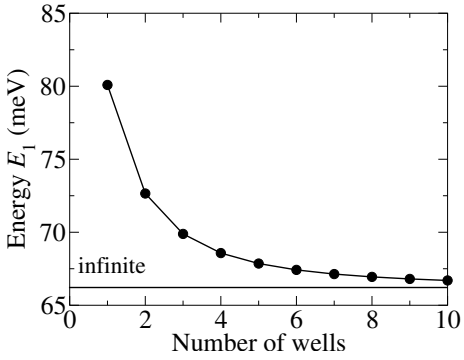


Figure 3.19: The ground state energy of an N-well heterostructure

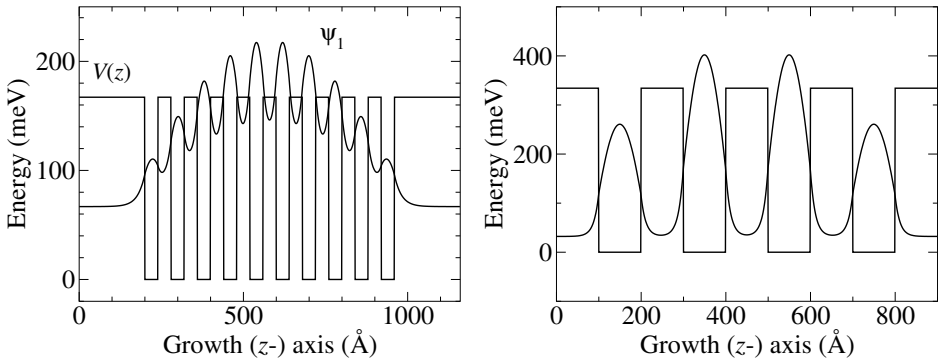


Figure 3.20: The ground state wave function of (a) a finite superlattice and (b) a multiple quantum well system

In the case of hydrogen, both electrons move into the lower-energy orbital, thus reducing the total energy of the two-hydrogen-atom system to form a chemically bound hydrogen molecule. In the case of the double quantum well here, a similar situation occurs with the electron spins aligning in an 'anti-parallel' arrangement in order to satisfy the Pauli exclusion principle.

Figure 3.18 displays the wave functions of the double quantum well with a central barrier width of 40 Å. Clearly, they form a symmetric and anti-symmetric pair, with the former being of lower energy, as discussed for the hydrogen molecule as above.

3.15 Multiple quantum wells and finite superlattices

Real layered structures can only contain a finite number of quantum wells, so infinite superlattices do not actually exist. Thus, the Kronig–Penney model derived in the previous chapter has to be used carefully, i.e. it will give the energy and dispersion relationships of carriers *near the centre of a many-period quantum well system.*

One way of discovering exactly how many quantum wells are required before a finite structure resembles an infinite would be to look at the ground state energy as a function of the number of periods. Figure 3.19 does exactly this for N repeats of a $40\,$Å GaAs/$40\,$Å $Ga_{0.8}Al_{0.2}As$ unit cell capped with $200\,$Å $Ga_{0.8}Al_{0.2}As$ barriers. By the time there are ten wells, the ground state energy is within $1\,\mathrm{meV}$ of that given by the Kronig–Penney model of an infinite structure.

While ten wells are sufficient from an energy perspective for this particular finite system to resemble the infinite case, the picture given by the wave function is quite different, as illustrated in Fig. 3.20(a). Clearly, this system is a superlattice in that the wave function 'associated' with each well has a significant overlap with that of the adjacent well. In addition, the finiteness of the structure leads to the feature that the electron is not equally likely to be in any of the wells. Compare this with the wave function for the heterostructure based on a $100\,$Å GaAs/$100\,$Å $Ga_{0.6}Al_{0.4}As$ unit cell, as in Fig. 3.20(b), which is clearly a multiple quantum well in that the wave function reaches zero between the wells. It should be noted, however, that the effect of finiteness is *still* apparent in that the probability of the electron being in any of the quantum wells is not equal.

3.16 Addition of electric field

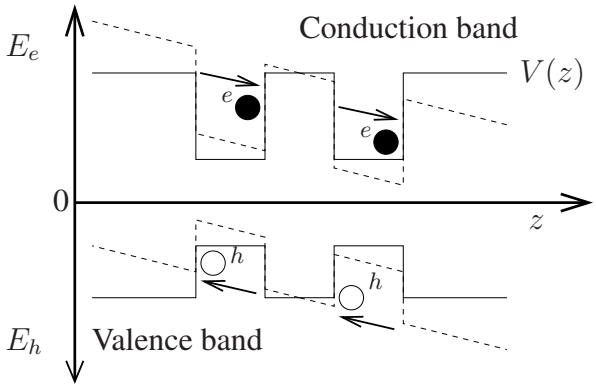


Figure 3.21: Schematic illustration of the tilting of the conduction and valence bands under the influence of an electric field

Unlike the previous analytical forms for the solution of Schrödinger's equation (described in Chapter 2), the numerical solution which includes the effect of an electric field on a

heterostructure is simplicity itself. The potential energy is simply added to the potential term within the appropriate shooting equation (equation (3.21) or (3.72)), i.e.

$$V(z) \rightarrow V(z) + qF(z - z_0) \tag{3.74}$$

where for an electron, $q = -e$ and for a hole, $q = +e$. The position z_0 represents the origin of the field, often chosen to be the centre of the well. Figure 3.21 shows the effect of an electric field on both the conduction- and valence-band potentials; note that increasing hole energies are measured downwards. Thus, any electrons in the double well are pulled to the right-hand side, while holes are pulled to the left, thus producing space charge or a polarisation of the carriers.

3.17 Extension to include variable permittivity

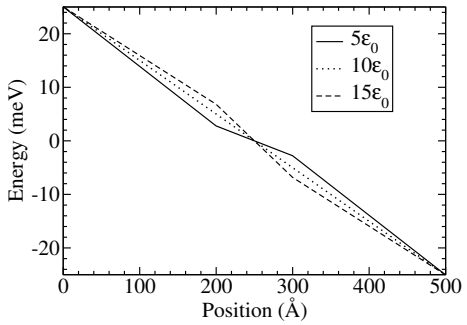


Figure 3.22: The shift in band-edge potential induced by an external $10\,\mathrm{kV\,cm^{-1}}$ electric field, for a $100\,$Å well with $\epsilon = 10\epsilon_0$, surrounded by $200\,$Å barriers with a range of permittivities

As was the case in Section 3.13, it is important to note that heterostructures, by definition, consist of materials with spatially varying properties. A spatially varying effective mass was introduced to modify the free-electron model within a heterostructure. Similarly, a spatially varying *relative permittivity* can be introduced to describe the electrostatic fields in a heterostructure.

Gauss's law states that:

$$\nabla[\epsilon(\mathbf{r})\mathbf{E}(\mathbf{r})] = -\rho(\mathbf{r}) \tag{3.75}$$

where $\epsilon(\mathbf{r}) = \epsilon_r(\mathbf{r})\epsilon_0$ is the permittivity at any point in space, $\mathbf{E}$ is the electric field and ρ is the charge density. For now, consider a structure that contains no charge. Recall also that the electric field is the derivative of potential $V(\mathbf{r})$. The equation, therefore, becomes:

$$\nabla\left[\epsilon(\mathbf{r})\nabla V(\mathbf{r})\right] = 0 \tag{3.76}$$

If the field is in the direction of growth, equation (3.76) reduces to one dimension and it can be converted into a discrete form, using the finite-difference approach. First, taking the inner derivative:

$$\frac{\partial}{\partial z}\left[\epsilon_i \frac{V_{i+1} - V_{i-1}}{2\delta z}\right] = 0 \tag{3.77}$$

and then the outer derivative:

$$\frac{1}{2\delta z}\left[\epsilon_{i+1}\frac{V_{i+2} - V_i}{2\delta z} - \epsilon_{i-1}\frac{V_i - V_{i-2}}{2\delta z}\right] = 0 \tag{3.78}$$

and finally, making the transformation $2\delta z \rightarrow \delta z$, the equation can be rearranged into the discretised form:

$$a_i V_{i-1} + b_i V_i + c_i V_{i+1} = 0 \tag{3.79}$$

where

$$a_{i+1} = c_i = \frac{\epsilon_{i+\frac{1}{2}}}{(\delta z)^2} \quad \text{and} \quad b_i = -\frac{\epsilon_{i+\frac{1}{2}} + \epsilon_{i-\frac{1}{2}}}{(\delta z)^2} \tag{3.80}$$

The boundary conditions for this expression are taken by 'pinning' the ends of the structure according to the external field, such that $V_0 = 0$ and $V_{N+1} = -eF(N+1)\delta z$. This leads to the system of linear equations:

$$\begin{aligned} b_1 V_1 + c_1 V_2 &= 0 \\ a_2 V_1 + b_2 V_2 + c_2 V_3 &= 0 \\ &\cdots \\ a_{N-1} V_{N-2} + b_{N-1} V_{N-1} + c_{N-1} V_N &= 0 \\ a_N V_{N-1} + b_N V_N &= c_N eF(N+1)\delta z \end{aligned} \tag{3.81}$$

Following the same approach as in Section 3.9, the expression can be represented succinctly in matrix form:

$$\mathsf{AV} = \boldsymbol{\rho} \tag{3.82}$$

where A is a symmetric tridiagonal matrix with the same form as equation (3.44), V is a column vector holding the samples of the field-induced potential and $\boldsymbol{\rho}$ contains all the right-hand-side values for the linear equations such that

$$\boldsymbol{\rho}^{\mathrm{T}} = \{0, 0, \ldots, c_N eF(N+1)\delta z\} \tag{3.83}$$

Whereas equation (3.74) simply gives a constant field across the entire structure, this modified calculation causes the field to be lower in regions with higher permittivities and vice versa. Figure 3.22 shows the combined effect on the band-edge potential, when a quantum well is surrounded by barriers with different permittivities. Carriers within a well surrounded by lower-permittivity barriers experience a reduced influence from external fields, since the *internal* potential gradient is reduced in the well region. The converse applies in systems in which the barrier permittivity is greater than that of the wells (as is the case in GaAs/GaAlAs systems).

3.18 Quantum-confined Stark effect

Section 2.10 demonstrated through perturbation theory the well-known phenomenon of the quantum-confined Stark effect. Figure 3.23 illustrates this energy level suppression as a function of field for the ground state of a single quantum well.

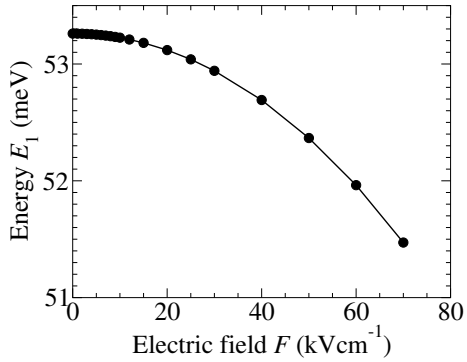


Figure 3.23: The effect of an electric field on the ground state energy E_1 of a 200 Å $Ga_{0.8}Al_{0.2}As$/60 Å GaAs/200 Å $Ga_{0.8}Al_{0.2}As$ single quantum well

In order to validate the perturbation theory approach of earlier, the curve represents a parabolic fit to the calculated data which are displayed as points on the figure. It was found that the ground state energy over the range of fields investigated could be represented very well by the parabola:

$$E_1(F) = E_1(0) - 0.00036F^2 \tag{3.84}$$

where $E_1(0)$ refers to the ground state energy ($53.310\,\mathrm{meV}$) at zero field, and the electric field strength F is in units of kilovolts per centimetre.

3.19 Field-induced anti-crossings

As a further example of electric-field-induced phenomena and the versatility of the numerical solution, consider a double quantum well, as shown in Fig. 3.16 but with the left-hand well wider than the right. This asymmetry breaks the degeneracy of the two confined states and results in the lower-energy state being more localised in the wider left-hand well and the higher-energy state being associated with the narrower right-hand well.

On the addition of an electric field, the left-hand side of the structure increases in potential, while the right-hand side decreases, and thus the energy levels are brought closer together. *A priori*, it is not clear what happens in such a situation. However, the system can be solved numerically for all field points in order to reveal the particulars of the quantum mechanics. Figure 3.24 displays the results of such a calculation.

The two energy levels *are* brought closer together, and as the field increases, their separation $\Delta E\,(= E_2 - E_1)$, which is initially decreasing linearly, reaches a minimum value

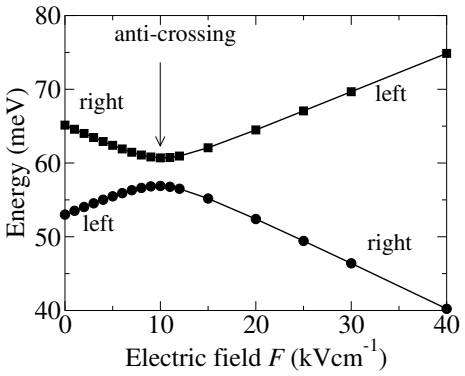


Figure 3.24: The two lowest-energy confined states of a 200 Å $Ga_{0.8}Al_{0.2}As$/60 Å GaAs/60 Å $Ga_{0.8}Al_{0.2}As$/50 Å GaAs/200 Å $Ga_{0.8}Al_{0.2}As$ double quantum well as a function of the applied electric field

and then increases again. Within the linear sections, ΔE is merely equal to the difference in field potentials between the centres of the wells, in this case:

$$\Delta E = eF\Delta z = eF(30\ \text{Å} + 30\ \text{Å} + 25\ \text{Å}) \tag{3.85}$$

At the minimum separation, the wave functions reveal a change in character—the lowest-energy state switches from being localised in the left-hand well to the right-hand well, and vice versa for the higher energy level. This phenomenon is referred to as an *anti-crossing*.

3.20 Symmetry and selection rules

Overlap integrals can be a very quick way of understanding or explaining complex interactions and are often instrumental in determining selection rules for carrier scattering events. The simplest selection rule is given by the orthonormality equation (3.46). As will be encountered later in Chapter 10, the relative strength of an *interband* electron–hole recombination (both excitonic and free-carrier) is proportional to O, even though one of the eigenstates corresponds to a conduction-band electron and the other to a valence-band hole; under the envelope function approximation, they are eigenstates of different Hamiltonians. For symmetric potentials, such as a single quantum well, the states have exact parity, and thus O is non-zero, so corresponding to an allowed transition, for $\Delta n = 0$, 2, etc.

Electron–electron scattering will also be discussed in Chapter 10; this is a scattering event between two initial states, resulting in two new states. The rate therefore contains an overlap between all four wave functions, as follows:

$$\text{rate} \propto \int_{-\infty}^{+\infty} \psi_i^* \psi_j^*\ \psi_f \psi_g\ \mathrm{d}z \tag{3.86}$$

Again, for electrons in symmetric potentials, the wave functions have definite parity, and thus for a two-subband system the $|2\rangle\,|2\rangle \rightarrow |1\rangle\,|1\rangle$ transition is allowed, while the $|2\rangle\,|2\rangle \rightarrow |2\rangle\,|1\rangle$ event is not.

Occasionally it is worthwhile to define another form for the overlap integral, i.e. the overlap integral of the moduli:

$$O_{\text{mod}} = \langle |\psi_n| \, | \, |\psi_{n'}| \rangle = \int_{-\infty}^{+\infty} |\psi_n^*| \, |\psi_{n'}| \, \mathrm{d}z \tag{3.87}$$

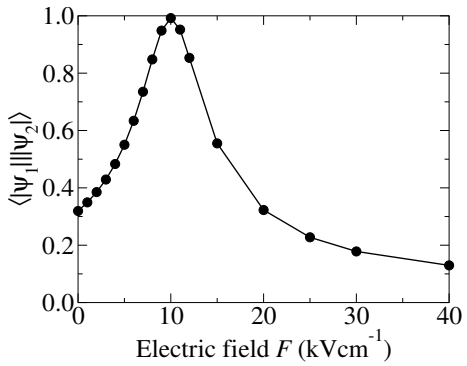


Figure 3.25: The overlap integral of the modulus of the wave functions undergoing the anti-crossing in Fig. 3.24

For example, Fig. 3.25 displays the overlap integral of the wave function moduli for the double quantum well of the previous section. As discussed there, the eigenstates undergo an anti-crossing at a field of $10\,\mathrm{kV\,cm^{-1}}$. Initially, the wave functions are localised in separate wells, and then as the field is increased the energy levels are brought into alignment and the wave functions 'spread out' over both wells. As both states are simultaneously eigenstates of the same Hamiltonian, O remains zero; however, the nature of the anti-crossing is reflected in the peak of O_{mod}.

3.21 The Heisenberg uncertainty principle

This is an opportune moment to digress slightly and use the techniques now available to explore a fundamental building block of quantum mechanics, namely Heisenberg's uncertainty principle. In particular, the aspect of interest in this work will be whether the eigenstates of one-dimensional finite potentials always obey the principle and how close to the limit they can be taken.

Mathematically, the uncertainty principle is given by the following [12]:

$$\Delta x_i \Delta p_j \gtrsim \frac{\hbar}{2} \delta_{ij} \tag{3.88}$$

where the δ function implies that for an uncertainty in the determination of the particle's position along the growth (z-)axis, there is no physical limit to the accuracy of any determination in the particle's in-plane momentum.

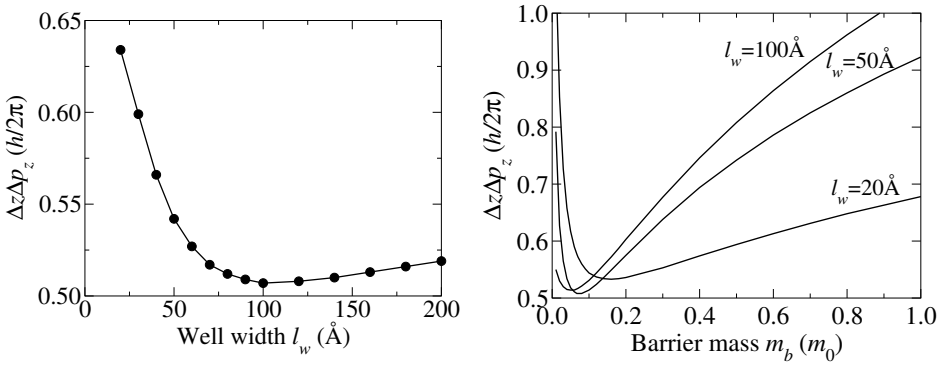


Figure 3.26: The product of the uncertainties in the position and momentum for an electron in a GaAs single quantum well, (a) surrounded by barriers of $100\,\mathrm{meV}$ height and with constant mass, as a function of the well width and (b) surrounded by $Ga_{0.6}Al_{0.4}As$ barriers, as a function of barrier mass and for various well widths

Restricting interest to the direction of confinement, the uncertainty relationship simplifies to:

$$\Delta z \Delta p_z \gtrsim \frac{\hbar}{2} \tag{3.89}$$

where

$$(\Delta z)^2 = \langle z^2 \rangle - \langle z \rangle^2 \qquad \text{and} \qquad (\Delta p_z)^2 = \langle p_z^2 \rangle - \langle p_z \rangle^2 \tag{3.90}$$

Section 2.7.3 discussed the relevance of Heisenberg's uncertainty relationship with respect to solutions to Schrödinger's equation in the limit of infinite barrier mass. At that point, the techniques had not been developed—the analytical forms for the wave functions do not lend themselves well to the calculation of Δz and Δp_z. Now, however, numerical solutions are at hand and the calculation can proceed.

The individual components contributing to Δz and Δp_z are expectation values of the state of interest, calculated in the usual way:

$$\langle z \rangle = \int_{-\infty}^{+\infty} \psi^*(z)\; z\; \psi(z)\; \mathrm{d}z \tag{3.91}$$

$$\langle z^2 \rangle = \int_{-\infty}^{+\infty} \psi^*(z)\; z^2\; \psi(z)\; \mathrm{d}z \tag{3.92}$$

$$\langle p_z \rangle = -i\hbar \int_{-\infty}^{+\infty} \psi^*(z)\; \frac{\partial}{\partial z}\psi(z)\; \mathrm{d}z \tag{3.93}$$

$$\langle p_z^2 \rangle = -\hbar^2 \int_{-\infty}^{+\infty} \psi^*(z)\; \frac{\partial^2}{\partial z^2}\psi(z)\; \mathrm{d}z \tag{3.94}$$

Whereas analytical calculations of $\langle p_z \rangle$ and $\langle p_z^2 \rangle$ are tedious, with the numerical formalism and the familiarity built up with finite-difference expansions, both quantities follow simply as:

$$\langle p_z \rangle = -i\hbar \int_{-\infty}^{+\infty} \psi^*(z) \frac{\psi(z+\delta z) - \psi(z-\delta z)}{2\delta z} \, \mathrm{d}z \tag{3.95}$$

and

$$\langle p_z^2 \rangle = -\hbar^2 \int_{-\infty}^{+\infty} \psi^*(z) \frac{\psi(z+\delta z) - 2\psi(z) + \psi(z-\delta z)}{(\delta z)^2} \, \mathrm{d}z \tag{3.96}$$

Figure 3.26(a) plots $\Delta z \Delta p_z$ for a typical single quantum well, as a function of the well width. The form of the curve is very interesting in that it is a non-monotonic function with a minimum at around $l_\mathrm{w} = 100\,\text{Å}$. This can be understood intuitively in terms of the spatial coordinate—as the well width is reduced from large values, Δz, which might be considered as the well width, decreases. However, at the same time the confinement energy increases and at narrow well widths lies just below the top of the well. At this point, the wave function 'spills over' the barriers, thus leading to an increase in Δz.

Interestingly, the minimum in $\Delta z \Delta p_z$ is just above the $\hbar/2$ limit given by Heisenberg's uncertainty principle. This is perhaps even more dramatically illustrated by the curves shown in Fig. 3.26(b), which plots $\Delta z \Delta p_z$ versus the barrier mass for a variety of well widths. These data correspond to the single quantum well shown in Fig. 2.22, which at the point of introduction was thought to imply a violation of Heisenberg's uncertainty principle, but now the calculations reveal this not to be the case.

The non-monotonic behaviour as a function of well width is again apparent from the ordering of the curves for any particular mass. In addition, a non-monotonic behaviour as a function of the electron mass in the barrier can be seen. The minimum value of $\Delta z \Delta p_z$ on Fig. 3.26(b) occurs for $l_\mathrm{w} = 50\,\text{Å}$, as would be expected from earlier. This occurs for a barrier mass which is very close to the well mass of $0.067 m_0$ and is equal to $0.508\hbar$, which is very close to but still larger than $\hbar/2$.

3.22 Extension to include band non-parabolicity

Section 1.8 introduced the concept of band non-parabolicity, in which an energy-dependent effective mass was used to describe the dispersion of carriers at energies above the band edge. This non-parabolic mass model is applicable for systems in which carriers are forced into high-energy states, for example, in narrow quantum wells. Using the energy-dependent effective mass from equation (1.32), the Schrödinger equation then becomes:

$$-\frac{\hbar^2}{2} \frac{\partial}{\partial z} \frac{1}{m^*(z,E)} \frac{\partial}{\partial z} \psi(z) + V(z)\psi(z) = E\psi(z) \tag{3.97}$$

which can be solved by using the iterative shooting equation (3.72) as before, but with the additional feature of adjusting the effective mass for each energy E.

Figure 3.27(a) displays the effect of the inclusion of non-parabolicity on the ground state energy of a single GaAs quantum well surrounded by $\mathrm{Ga}_{1-x}\mathrm{Al}_x\mathrm{As}$ barriers. This figure plots ΔE, the energy calculated including non-parabolicity minus the energy not including non-parabolicity. Clearly, non-parabolicity increases the energy calculated, and the increase is

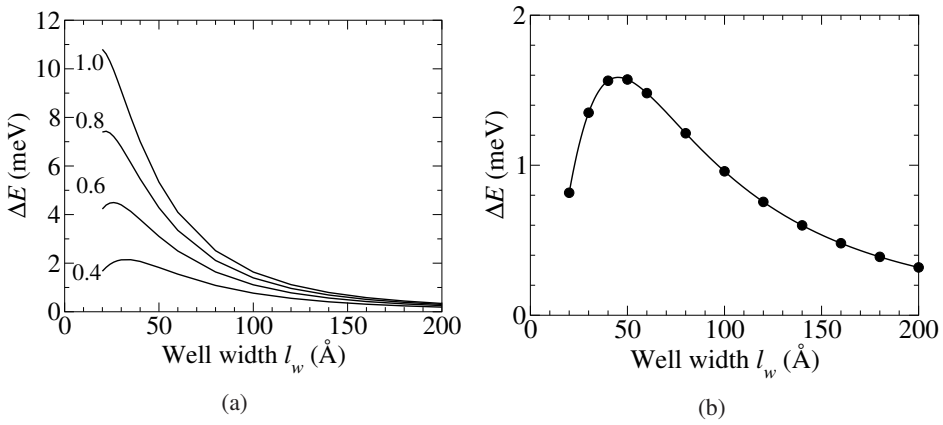


Figure 3.27: The difference in ground state energy of an electron at the subband minimum in a single quantum well, with and without non-parabolicity, $\Delta E = E_1(\text{with}) - E_1(\text{without})$, for (a) a GaAs/GaAlAs quantum well with a range of $Ga_{1-x}Al_xAs$ barrier concentrations x and (b) an $In_{0.53}Ga_{0.47}As/In_{0.52}Al_{0.48}As$ well

(generally) larger for an increasing barrier height and a decreasing well width. The latter two points reflect the expected conclusions formed earlier, i.e. that as the electron (or hole) is forced up the E–**k** curve, the parabolic description of the band becomes more approximate.

The barrier material, $Ga_{1-x}Al_xAs$, becomes indirect as x increases beyond 0.42, thus producing an extra complication. Many real layered structures are grown with concentrations below this value, and therefore the majority of examples in this book do conform to this criterion. Given this, the *maximum* effect of band non-parabolicity is largely given by the bottom curve of Fig. 3.27(a), i.e. $2\,\text{meV}$. The chapters that follow are centred around the development of techniques and methods and the illustration of generic physics which is applicable across material systems, so band non-parabolicity will not be employed in future examples of calculations based on the $GaAs/Ga_{1-x}Al_xAs$ material system.

One material in which this effect is thought to be more important is the quaternary $In_{1-x-y}Al_xGa_yAs$ system, which has become an important material system for intersubband lasers (more of this later in Chapter 10) because of its large conduction-band offset. In particular, the $In_{0.52}Al_{0.48}As$ and $In_{0.53}Ga_{0.47}As$ alloys are lattice-matched to the readily available InP substrates (see [13] p. 196).

In fact, Fig. 3.27(b) demonstrates that the effect is relatively small, but nevertheless significant, at narrower well widths. Indeed, energy level spacings in intersubband lasers are designed in relation to phonon energies which are typically a few tens of millielectronvolts, and hence it is debatable whether a $2\,\text{meV}$ effect should be included. The effect is larger when the $In_{0.53}Ga_{0.47}As$ wells are surrounded by strained AlAs barriers, so providing even more offset, and this material system has been the subject of several recent papers [14, 15, 16].

Incorporating band non-parabolicity into matrix methods is, unfortunately, more complicated than with the shooting method, because the energy appears within the

coefficients in the H matrix (equation (3.44)). This means that the matrix no longer contains constant coefficients, and is instead a function of energy. As such, equation (3.43) becomes a *nonlinear* eigenvalue problem, which cannot be solved directly using the rapid methods described in Section 3.9.

The simplest matrix method is to solve the problem iteratively by splitting the energy range into cells (as with the shooting method) and attempt to find the eigenvalues of equation (3.43) using:

$$\det\left[E\mathrm{I} - \mathsf{H}(E)\right] = 0 \tag{3.98}$$

where I is the identity matrix of the same order as the Hamiltonian matrix. However, as with the shooting method, this is a slow process that relies on an appropriate choice of energy cell size. More sophisticated numerical methods have been developed, which allow solutions to be located directly, for example by converting the equation into a linear (but much larger) eigenvalue problem [17][2] or by solving for the conduction- and valence-band states simultaneously [18]. In all cases, however, these direct solutions are beyond the scope of this book, and may be much slower than iterative solutions (although highly accurate).

3.23 Poisson's equation

All of the techniques are now in place to be able to solve the Schrödinger equation for any heterostructure for which the band-edge potential profile defining the structure is known. However, all of the theoretical methods and examples described so far have concentrated solely on solving systems for a single charge carrier. In many devices such models would be inadequate as large concentrations of charge carriers can be present, owing to doping in the material. For example, free electrons are introduced into the conduction band when donor atoms are added to a crystal. As before, the Hamiltonian for this system is found by considering all the forces acting on an individual electron (or hole). The regions with high electron density are repulsive to an electron, owing to the negative charge, while the regions containing ionised donors are attractive to an electron, owing to the positive charge. These space-charge effects can be represented by an additional potential term in the Hamiltonian. In order to decide whether or not typical carrier densities would give rise to a *significant* additional potential on top of the usual band-edge potential terms (which will be labelled specifically as V_{CB} or V_{VB}), it then becomes necessary to solve the electrostatics describing the system.

When considering the case of an n-type material, it is worth stating (although obvious) that the number of 'free' electrons in the conduction band is equal to the number of positively charged ionised donors in the heterostructure. In reality, the concentration of *ionised* donors (and hence free carriers) will differ somewhat from the total concentration of donor atoms in the material, since a small is needed to excite carriers into the relevant band of the crystal. However, this adds significant complexity to calculations and can be ignored in many common circumstances. In the example shown in Fig. 3.28 this equates to six donors becoming ionised and supplying six electrons into the quantum well—thus the system maintains charge neutrality. Figure 3.28 is a *modulation*-doped system in that the free carriers become spatially separated from the ions that supply them.

[2]This method is included in the QWWAD simulation software tools.

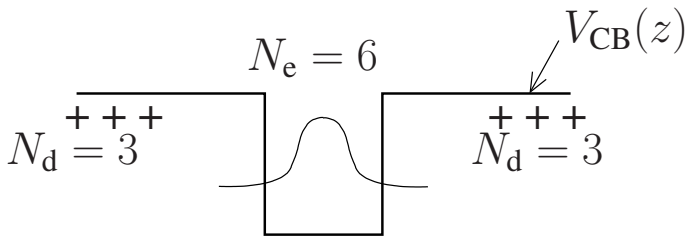


Figure 3.28: A modulation-doped single quantum well

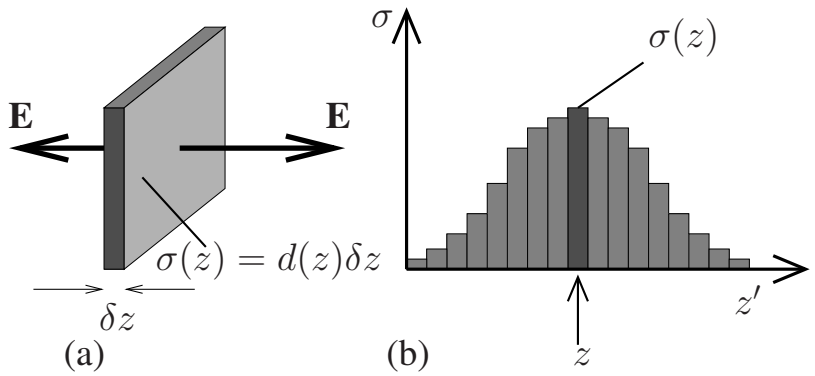


Figure 3.29: Electric field strength from an infinite plane of charge of volume density $d(z)$ and thickness δz

The additional potential term $V_\rho(z)$ arising from this, or any other charge distribution ρ, can be expressed using Poisson's equation:

$$\nabla^2 V_\rho = -\frac{\rho}{\epsilon} \tag{3.99}$$

where ϵ is the permittivity of the material, $\epsilon = \epsilon_r \epsilon_0$. The solution is generally obtained via the electric field strength $\mathbf{E}$. Recalling that:

$$\mathbf{E} = -\nabla V \tag{3.100}$$

the potential then would follow in the usual way [19]:

$$V_\rho(\mathbf{r}) = -\int_{-\infty}^{\mathbf{r}} \mathbf{E}_\bullet \mathrm{d}\mathbf{r} \tag{3.101}$$

Given that the potential profiles, $V_{\mathrm{CB}}(z)$ for example, are one-dimensional, then they will also produce a one-dimensional charge distribution. In addition, remembering that the quantum wells are assumed infinite in the x–y plane then any charge density $\rho(z)$ can be

thought of as an infinite plane, i.e. a *sheet*, with areal charge density $\sigma(z)$ and thickness δz, as shown in Fig. 3.29(a). Such an infinite plane of charge produces an electric field perpendicular to it, and with a strength:

$$\mathbf{E} = \frac{\sigma}{2\epsilon} \tag{3.102}$$

Note that, as the sheet is infinite in the plane, the field strength is constant *for all distances from the plane*. The total electric field strength due to many of these planes of charge, as shown in Fig. 3.29(b), is then the sum of the individual contributions as follows:

$$\mathbf{E}(z) = \sum_{z'=-\infty}^{\infty} \frac{\sigma(z')}{2\epsilon} \text{sign}(z - z') \tag{3.103}$$

where the function sign is defined as

$$\text{sign}(z) = \begin{cases} +1, & z \geq 0 \\ -1, & z < 0 \end{cases} \tag{3.104}$$

and has been introduced to account for the vector nature of $\mathbf{E}$, i.e. if a single sheet of charge is at a position z', then for $z > z'$, $\mathbf{E}(z) = +\sigma/(2\epsilon)$, whereas for $z < z'$, $\mathbf{E}(z) = -\sigma/(2\epsilon)$. Note further that it is only the charge neutrality, i.e. there are as many ionised donors (or acceptors) in the system as there are electrons (or holes), or, expressed mathematically:

$$\sum_{z=-\infty}^{+\infty} \sigma(z) = 0 \tag{3.105}$$

which ensures that the electric field, and hence the potential, go to zero at large distances from the charge distribution. For the case of a doped semiconductor, there would be two contributions to the charge density $\sigma(z)$, where the first would be the ionised impurities and the second the free charge carriers themselves. While the former would be known from the doping density in each semiconductor layer, as defined at growth time, the latter would be calculated from the probability distributions of the carriers in the heterostructure. Thus if $d(z)$ defines the volume density of the dopants at position z, where the planes are separated by the usual step length δz, then the total number of carriers, per unit cross-sectional area, introduced into the heterostructure is given by:

$$N = \int_{-\infty}^{+\infty} d(z)\, \mathrm{d}z \tag{3.106}$$

The net charge density in any of the planes follows as:

$$\sigma(z) = q\left[N\psi^*(z)\psi(z) - d(z)\right]\delta z \tag{3.107}$$

where q is the charge on the extrinsic carriers. The step length δz selects the proportion of the carriers that are within that 'slab' and converts the volume density of dopant, $d(z)$, into an areal density.

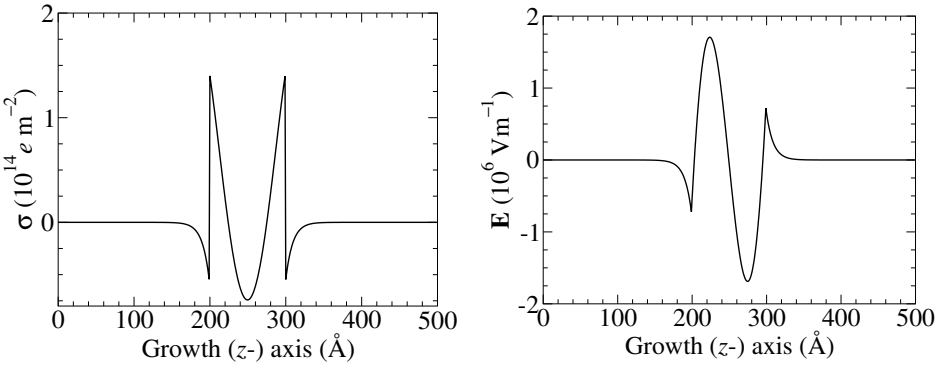


Figure 3.30: (a) Areal charge density σ and (b) electric field strength $\mathbf{E}$ for a 100 Å GaAs well, n-type doped to $2 \times 10^{18}\,\mathrm{cm}^{-3}$, surrounded by undoped $Ga_{0.8}Al_{0.2}As$ barriers

If the charge carriers are distributed over more than one subband, then the contribution to the charge density $\sigma(z)$ would have to be summed over the relevant subbands, i.e.

$$\sigma(z) = q\left(\sum_{i=1}^{n} N_i \psi_i^*(z)\psi_i(z) - d(z)\right)\delta z \tag{3.108}$$

where $\sum_{i=1}^{n} N_i = N$.

Figure 3.30(a) shows the areal charge density along the growth axis for a 100 Å GaAs well, n-type doped to $2 \times 10^{18}\,\mathrm{cm}^{-3}$, surrounded by 200 Å undoped $Ga_{0.8}Al_{0.2}As$ barriers. The ionised donors yield a constant contribution to σ within the well of $d(z)\delta z = 2 \times 10^{24}\,\mathrm{m}^{-3} \times 1\,\text{Å} = +2 \times 10^{14}\,\mathrm{m}^{-2}$, in each of the 1 Å thick slabs. Hence, the total number N of electrons in the quantum well is $100 \times 2 \times 10^{14}\,\mathrm{m}^{-2} = 2 \times 10^{12}\,\mathrm{cm}^{-2}$. By assuming that the electrons introduced by such doping all occupy the ground state of the quantum well, the curve on top of the ionised impurity background clearly resembles $-\psi^*\psi$, as expected from the mathematics. The discontinuities in σ occur at the edges of the doping profiles and *are* of magnitude $2 \times 10^{14}\,\mathrm{m}^{-2}$, again as expected.

There are a number of points to note about Fig. 3.30(b), which plots the electric field strength $\mathbf{E}$ due to the charge distribution (as in equation (3.103)) along the growth axis of the heterostructure. First, the field reaches zero at either end of the structure, which implies charge neutrality. In addition, the zero field at the centre of the structure reflects the symmetry of the charge distribution. The electric field strength itself is not an observable, merely an intermediate quantity which *can* be useful to plot from time to time; the quantity which is significant is, of course, the potential due to this charge distribution. Figure 3.31 plots the potential as calculated from equation (3.101), as usual defining the origin, in this case for the potential, at the effective infinity μ at the left-hand edge of the barrier–well–barrier structure.

Again, the symmetry of the original heterostructure and doping profiles is reflected in the symmetric potential. The potential is positive at the centre of the well since the system under

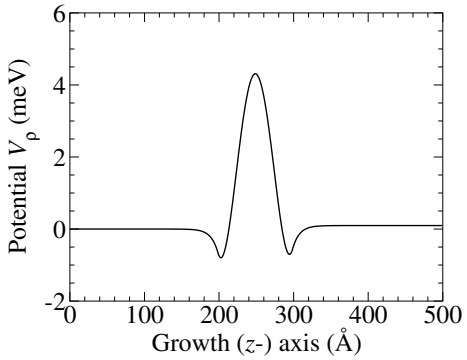


Figure 3.31: The potential due to the ionised donor/electron charge distribution

consideration consists of electrons in the conduction band, so any test charge used to probe the potential is also an electron which would be repelled by the existing charge. The carrier density in this single quantum well is reasonably high at $2 \times 10^{12}\,\mathrm{cm}^{-2}$, and this produces a potential of up to $4\,\mathrm{meV}$; while this is small compared to the conduction-band offset, which is usually of the order of 100 or $200\,\mathrm{meV}$ or more, it could still have a measurable effect on the energy eigenvalues of the quantum well.

3.24 Matrix solution of Poisson's equation

Poisson's equation is a more general form of equation (3.76), which described the internal potential within an uncharged structure. The only necessary change is to account for the non-zero charge density on the right-hand side of the equation:

$$\nabla\left[\epsilon(\mathbf{r})\nabla V(\mathbf{r})\right] = -\rho(\mathbf{r}) \tag{3.109}$$

The same matrix solution is obtained as in Section 3.17, with the exception that the right-hand side vector now contains samples of the charge density at each point instead of zeros:

$$\boldsymbol{\rho}^{\mathrm{T}} = \{\rho_1, \rho_2, \ldots, \rho_N + c_N eF(N+1)\delta z\} \tag{3.110}$$

The resulting matrix equation can be solved using the same techniques as before. This is a particularly useful method, since it accounts both for internal space-charge effects and external field within the same equation. It is worth noting that only a fixed-potential boundary condition has been considered here, i.e. with the edges of the structure 'pinned' to a known external potential. A range of other boundary conditions may be appropriate in other circumstances, for example, if the structure is periodic, or if the electric field is set to zero.

3.25 Self-consistent Schrödinger–Poisson solution

The energy eigenvalues are found by considering the introduction of a further test electron into the system and incorporating the potential due to the carrier density already present into

the Schrödinger equation, i.e. the potential term $V(z)$ in equation (3.61) becomes:

$$V(z) \rightarrow V_{\mathrm{CB}}(z) + V_{\rho}(z) \tag{3.111}$$

where V_{CB} represents the band-edge potential at zero doping and V_{ρ} represents the potential due to the non-zero number of carriers, i.e. the charge density ρ.

A numerical Schrödinger equation solver, as described in detail earlier in this chapter, can be used without alteration to solve for this new potential, which will thus yield new energies and wave functions. The latter is an important point since the potential due to the charge distribution is itself dependent on the wave functions. Therefore, it is necessary to form a closed loop solving Schrödinger's equation, calculating the potential due to the resulting charge distribution, adding it to the original band-edge potential, solving Schrödinger's equation again, and so on—a process which is illustrated schematically in Fig. 3.32.

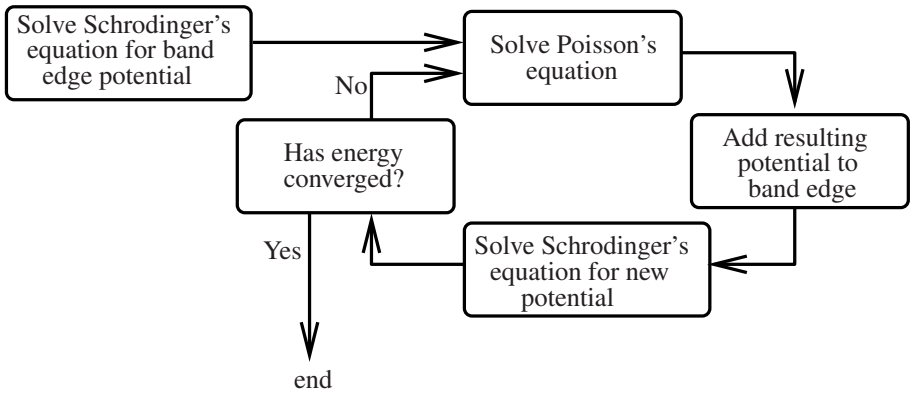


Figure 3.32: Block diagram illustrating the process of self-consistent iteration

The process is repeated until the energy eigenvalues converge; at this point the wave functions are simultaneously solutions to both Schrödinger's and Poisson's equations—the solutions are described as *self-consistent*, rather like Hartree's approach to solving many electron atoms (see [6], p. 396).

Figure 3.33 shows the result of adding the potential due to the charge distribution V_{ρ}, as displayed in Fig. 3.31, to the original band-edge potential V_{CB} for the single quantum well of the previous section. The perturbation, even at this relatively high carrier density of $2 \times 10^{12}\,\mathrm{cm}^{-2}$, is rather small compared with the barrier height, for instance. Nonetheless, it is important to calculate the effect of this perturbation on the electron energy levels by continuing with the iterative process and looking for convergence of the resulting energy solutions. This process is illustrated in Fig. 3.34.

The first iterative loop produces the greatest change in the energy level from the single carrier system to the doped system. Subsequent iterations produce only minor refinements to the energy level, which has certainly converged by the eighth loop. Altogether the ground state energy has changed by about $2\,\mathrm{meV}$ when accounting for this finite charge density.

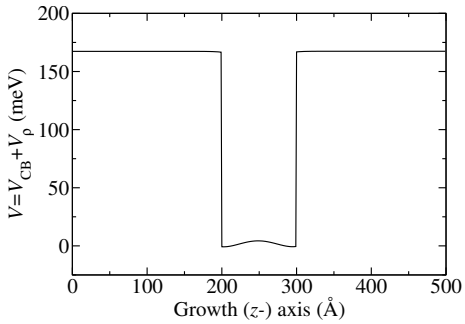


Figure 3.33: The sum of the band-edge potential V_{CB} and Poisson's potential V_ρ for the single quantum well of Fig. 3.30

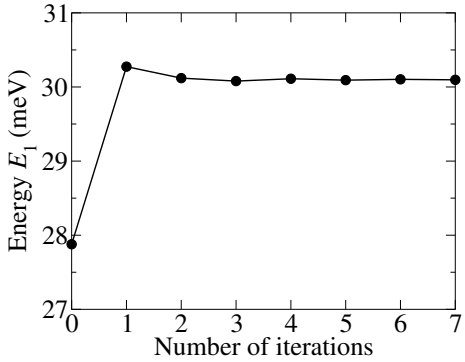


Figure 3.34: The ground state energy of the single quantum well of Fig. 3.30 as a function of the number of iterative steps

It is worth noting that numerical precision plays an important role in these self-consistent calculations. For example, small asymmetries in the calculated potential can cause unrealistic charge accumulation or depletion within simulated structures. In some cases, this can cause the iterative calculation of the Poisson and Schrödinger equations to become unstable, and to oscillate or diverge rather than to converge on a solution.[3] As such, it is generally important to record the energies of states at each iteration and to verify that the solution has indeed converged. This instability is particularly problematic in the shooting method approach, since numerical errors accumulate as the solution propagates across the structure, as illustrated in Section 3.7. By contrast, matrix methods are generally much more stable. Instability in self-consistent solutions can often be resolved by using a finer spatial sampling period, or by ensuring that subband energies are computed to high precision in each iteration.

[3]This was discussed in Section 3.19 of the third edition of this book.

3.26 Modulation doping

Although mention has been made of quantum well systems in which doping in the barriers leads to a spatial separation of the ions and charge carriers, i.e. which collect in a quantum well, quantitative calculations presented thus far have not considered these modulation-doped systems. Figure 3.35 shows the band-edge potential, V_{CB}, and the self-consistent potential, $V_{\mathrm{CB}} + V_{\rho}$, for a system of the type illustrated in Fig. 3.28, i.e. an undoped single quantum well surrounded by doped barriers, with the full layer definition thus being: 100 Å $Ga_{0.8}Al_{0.2}As$ doped n-type to $2 \times 10^{17}\,\mathrm{cm}^{-3}$; 100 Å GaAs undoped; 100 Å $Ga_{0.8}Al_{0.2}As$ doped n-type to $2 \times 10^{17}\,\mathrm{cm}^{-3}$.

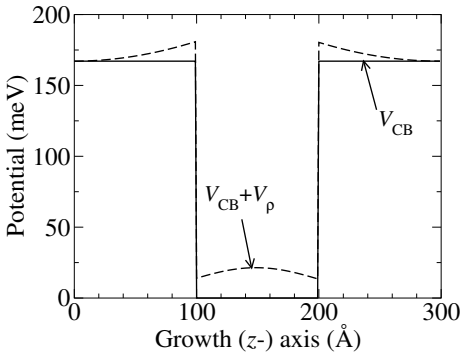


Figure 3.35: The band-edge potential and the self-consistent potential of a modulation-doped single quantum well

The electrons introduced into the system are physically separated from the ionised donors, so therefore instead of an ion/charge carrier plasma, the mobile charge in this case is often referred to as a *two-dimensional electron gas*. The physical separation leads to a reduction in the ionised impurity scattering and hence increased electron mobilities for in-plane (x–y) transport, a feature which is exploited in *high-electron-mobility transistors*.

3.27 The high-electron-mobility transistor

The high-electron-mobility transistor (HEMT), the heterostructure field-effect transistor (HFET), the modulated-doped field-effect transistor (MODFET), or even the two-dimensional electron-gas field-effect transistor (TEGFET), are all names for a transistor designed to exploit the high in-plane (x–y) mobility which arises when a (usually single) heterojunction is modulation-doped. In essence, a high-bandgap material such as $Ga_{1-x}Al_xAs$ is doped n-type, and upon ionisation of the donors the electrons move to the lower energy levels in the narrower-bandgap material, such as GaAs—this situation is very similar to the single well of the previous section, but with just one barrier layer in this case.

Figure 3.36(a) shows the evolution of the self-consistent potential, $V_{\mathrm{CB}} + V_{\rho}$, with each iteration. After the first solution of Schrödinger's equation, the potential V consists entirely

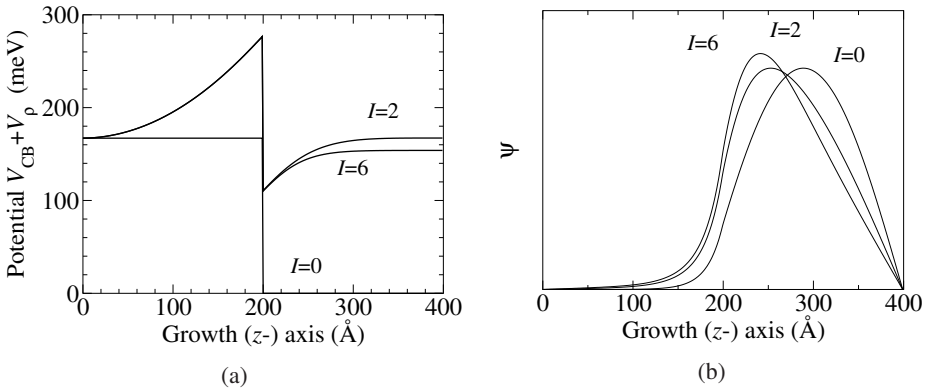


Figure 3.36: The self-consistent evolution of (a) the potential $V_{\mathrm{CB}} + V_\rho$ and (b) the ground state wave function, with the number of iterations I, for a single modulation-doped heterojunction

of the band-edge potential V_{CB} and clearly defines the single heterojunction. However, as the iteration proceeds the charge carriers are drawn towards the interface and form a triangular well potential—a process often referred to as *band bending*. This is vividly illustrated by Fig. 3.36(b), which shows the corresponding evolution of the electron wave functions.

This section has merely 'scratched the surface' of transistor modelling, and has been used to illustrate the applicability of quantum and electrostatic theory to everyday electronic devices, as well as the more esoteric semiconductor structures concentrated on in this work. For more information on heterojunction-based transistors, see Kelly [20], Chapter 16, or more specialised works, such as [21, 22, 23, 24].

3.28 Band filling

One way of measuring the energy levels of quantum well systems is via the use of spectroscopy, which involves illuminating the semiconductor with light from a laser. If the photon energy is greater than the band gap then, in the bulk material, electrons are excited from the valence-band states to the conduction band, which, depending in more detail on the excitation energy, can produce pairs of free electrons and free holes, or excitons (see Chapter 6). In a quantum well system, the photons must have additional energy in order to overcome the electron and hole ground state confinement energies as well.

In a doped system, e.g. n-type, many electrons already exist in the conduction band, and thus spectroscopic measurements will observe the self-consistent solutions, $E_{\mathrm{e1}}^{\mathrm{s.c.}}$ and $E_{\mathrm{h1}}^{\mathrm{s.c.}}$ for the confinement energies.

However, there is another effect which may or may not be large enough to observe. Thus far, it has been assumed that there are an unlimited number of states available to electrons (or holes) when they are introduced as extrinsic carriers; note (see Chapter 2) that the density of states has been derived for two-dimensional systems. Thus excitation can only occur when

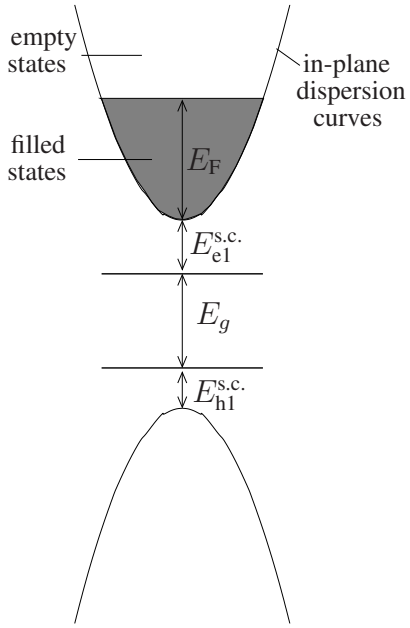


Figure 3.37: Schematic representation of the increase in band edge due to band filling

the photon energy is sufficiently large to excite an electron into an *empty* state (see Fig. 3.37). At low temperatures, the Fermi edge defines a sharp boundary between the filled states $f^{\mathrm{FD}}(E) = 1$ and the empty states $f^{\mathrm{FD}}(E) = 0$, and hence the *minimum* excitation energy must overcome this additional component, i.e.

$$E = E_{\mathrm{g}} + E_{\mathrm{e1}}^{\mathrm{s.c.}} + E_{\mathrm{h1}}^{\mathrm{s.c.}} + E_{\mathrm{F}} \tag{3.112}$$

This phenomenon of an apparently increased 'band edge' to absorption is known as *band filling*. This can also occur in undoped systems under very high excitation intensities which produce large numbers of both electrons and holes, which in turn can fill all of the available states and thus increase the minimum required photon energy for absorption. It can be accounted for under the present scheme by simply calculating the quasi-Fermi energy, E_{F}, as described in Section 2.5.

Exercises

(1). Consider the function $y = \ln\left[\sin(x^2) + 1.5\right]$.

(a) Locate the root of the function that lies in the range $\pi < x < \frac{3\pi}{2}$, using (i) an analytical derivation to obtain an exact value, expressed in terms of π, and (ii) a linear bisection algorithm to a precision of 3 decimal places.

(b) Calculate the first and second derivatives of the function at this root (i) analytically and (ii) numerically using a finite-difference approximation.

(c) Use the Newton–Raphson algorithm to locate roots of the function to a precision of 3 decimal places. Determine the approximate set of starting values x_0 for which the algorithm converges on the desired root.

(2). Use the `qwwad_ef_parabolic_well` and `qwwad_ef_band_edge` programs to generate the potential profile for a parabolic quantum well with 200 Å width and a maximum alloy composition of 100%. Use the `qwwad_ef_generic` program to compute the energies of all bound states using (a) parabolic bands and (b) non-parabolic bands. Determine the accuracy of the numerical solutions by comparison with the analytical expression, and explain any deviation qualitatively.

(3). Use the `qwwad_mesh` and `qwwad_ef_band_edge` programs to generate the potential profiles for a pair of 50 Å GaAs quantum wells, separated by a $Ga_{0.6}Al_{0.4}As$ barrier layer. Compute the first two bound states in the system using the `qwwad_ef_generic` program, for a range of barrier widths. Explain *quantitatively* the energies of the first two solutions in the limiting cases of extremely thick and extremely thin layers. (Hint: use the `qwwad_ef_square_well` program to compute the solutions for single wells of appropriate widths).

(4). A matrix with a constant value along each of its diagonal bands is known as a *Toeplitz matrix*. The eigenvalues of a tridiagonal Toeplitz matrix are given by $E_n = b - 2\sqrt{ac}\cos[n\pi/(N+1)]$, where N is the order of the matrix and a, b and c are the subdiagonal, diagonal and superdiagonal elements, respectively. Hence, solve equation (3.43) exactly for the energies of states in a constant potential and prove that the solutions of the discretised Schrödinger equation approach the analytical solutions of an infinite quantum well when the step size δz approaches zero.

(5). Use the `qwwad_ef_generic` program to compute the first 20 bound states in a structure containing ten 60 Å GaAs quantum wells, each separated by a 60 Å $Ga_{0.6}Al_{0.4}As$ barrier layer. Explain the computed band structure, with reference to the results in Section 2.11. Write a script to repeatedly adjust the well and barrier widths in this structure, and compare the energies of the first and tenth states with the analytical solutions in Fig. 2.30.

(6). Use the `qwwad_ef_generic` program to determine the separation between the lowest pair of subband energies in a double quantum well system. Analyse the variation in anti-crossing energy with respect to structural properties (e.g., barrier width, potential and well asymmetry).

(7). The first terahertz-frequency quantum cascade laser with peak output power greater than 1 W [25] consisted of a periodic system of four GaAs wells with $Ga_{0.84}Al_{0.16}As$ barriers in each period. The layer thicknesses are **52**/103/**17**/107.5/**36**/88/**39.5**/<u>172</u> Å, where the barrier thicknesses are stated in bold, wells are in regular type, and an n-doped well layer ($3 \times 10^{16}\,\mathrm{cm}^{-3}$) is underlined. The device operated in an external electric field of $7.6\,\mathrm{kV\,cm}^{-1}$.

(a) Select appropriate programs from the QWWAD software suite to simulate one period of the band structure, taking into account all internal and external electric field effects self-consistently. Compute the first five states in the system. (Hint:

embed the structure within a pair of thick barriers and use the `qwwad_poisson` program with the `--mixed` boundary conditions option to ensure electric field continuity at the ends of the structure).

(b) The optical transition takes place between the states that are tightly confined in the first and second quantum wells in the system. Compare the simulated transition energy with the experimentally measured $3.4\,\mathrm{THz}$ emission frequency.

(c) Electrons are extracted from the structure via a longitudinal optic phonon interaction (see Section 10.6). This process transfers electrons between the two lowest bound states in the last well in the system. Compare the simulated extraction-transition energy with the $36\,\mathrm{meV}$ phonon energy.

(d) Repeat the calculation using a non-parabolic band model. Comment on the accuracy and speed of the simulation, compared with the parabolic model. On balance, select which model would be most appropriate for a detailed simulation of the device (i.e. using hundreds or thousands of self-consistent calculations at different bias points and temperatures).

(8). Consider a single $100\,\text{Å}$ GaAs/$Ga_{0.6}Al_{0.4}As$ quantum well with a thin, highly doped film at the centre.

(a) Calculate the perturbation to the ground state energy self-consistently as the sheet doping density is varied between 1×10^9 and $1 \times 10^{10}\,\mathrm{cm}^{-2}$ in a $10\,\text{Å}$ wide doped layer.

(b) Adjust the width of the doped layer, while maintaining a constant doping density of $1 \times 10^{10}\,\mathrm{cm}^{-2}$. This provides a simple model of the effect of dopant atoms diffusing away from their intended location. Investigate the influence that this has on the ground state energy. Comment on the implications that this has for the practical development of heterostructure devices, in which unintended dopant diffusion may occur during fabrication.

References

[1] J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho, 'Quantum cascade laser', *Science*, **264**(5158):553–556, 1994.

[2] R. P. Brent, *Algorithms for Minimization without Derivatives*, Prentice Hall, Englewood Cliffs, N.J., 1973.

[3] M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, and F. Rossi, *GNU Scientific Library Reference Manual.*, Network Theory Limited, Bristol, Third (v1.12) edition, 2009.

[4] J. P. Killingbeck, *Microcomputer Algorithms*, Hilger, Bristol, 1992.

[5] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, Third edition, 1999.

[6] R. M. Eisberg, *Fundamentals of Modern Physics*, John Wiley & Sons, Inc., New York, 1961.

[7] D. K. Ferry, *Quantum Mechanics: An Introduction for Device Physicists and Electrical Engineers*, IOP Publishing, London, 1995.

[8] P. A. M. Dirac, *The Principles of Quantum Mechanics*, Clarendon Press, Oxford, Fourth edition, 1967.

[9] E. H. Li, 'Interdiffusion as a means of fabricating parabolic quantum wells for the enhancement of the nonlinear third-order susceptibility by triple resonance', *Appl. Phys. Lett.*, **69**:460–462, 1996.

[10] S. Flügge, 'One-body problems without spin,' in *Practical Quantum Mechanics*, number 177 in Classics in Mathematics, pp. 25–331. Springer, Berlin, 1999.

[11] S. R. Jackson, J. E. Nicholls, W. E. Hagston, P. Harrison, T. Stirner, J. H. C. Hogg, B. Lunn, and D. E. Ashenford, 'Magneto-optical study of exciton binding energies band offsets and the role of interface potentials in CdTe/$Cd_{1-x}Mn_xTe$ multiple quantum wells', *Phys. Rev. B*, **50**:5392–5403, 1994.

[12] F. Schwabl, *Quantenmechanik*, Springer, Berlin, 1990.

[13] H. Landolt and R. Börnstein, Eds., *Numerical Data and Functional Relationships in Science and Technology*, vol. 22a of *Series III*, Springer-Verlag, Berlin, 1987.

[14] P. Harrison and R. W. Kelsall, '1.55 μm intersubband pumping of an $In_{0.53}Ga_{0.47}As$/AlAs:InP symmetric double quantum well terahertz laser', *Physica E*, **2**(1–4):468–472, 1998.

[15] J. H. Smet, L. H. Peng, Y. Hirayama, and C. G. Fonstad, 'Electron intersubband transitions to 0.8 eV (1.55 μm) in InGaAs/AlAs single quantum wells', *Appl. Phys. Lett.*, **64**(8):986–987, 1994.

[16] Y. Hirayama, J. H. Smet, L.-H. Peng, C. G. Fonstad, and E. P. Ippen, 'Feasibility of 1.55 μm intersubband photonic devices using InGaAs/AlAs pseudomorphic quantum well structures', *Jpn. J. Appl. Phys.*, **33**(1S):890, 1994.

[17] J. D. Cooper, A. Valavanis, Z. Ikonić, P. Harrison, and J. E. Cunningham, 'Finite difference method for solving the Schrödinger equation with band nonparabolicity in mid-infrared quantum cascade lasers', *J. Appl. Phys.*, **108**(11):113109, 2010.

[18] X. Ma, K. Li, Z. Zhang, H. Hu, Q. Wang, X. Wei, and G. Song, 'Two-band finite difference method for the bandstructure calculation with nonparabolicity effects in quantum cascade lasers', *J. Appl. Phys.*, **114**(6):063101, 2013.

[19] W. J. Duffin, *Electricity and Magnetism*, McGraw-Hill, London, Third edition, 1980.

[20] M. J. Kelly, *Low Dimensional Semiconductors: Materials, Physics, Technology, Devices*, Clarendon Press, Oxford, 1995.

[21] C. M. Snowden, Ed., *Semiconductor Device Modelling*, Springer-Verlag, London, 1988.

[22] C. M. Snowden and R. E. Miles, Eds., *Compound Semiconductor Device Modelling*, Springer-Verlag, London, 1993.

[23] P. H. Ladbrooke, *MMIC Design: GaAs FETs and HEMTs*, Artech House, Boston, 1989.

[24] J. M. Golio, *Microwave MESFETs and HEMTs*, Artech House, Boston, 1991.

[25] L. Li, L. Chen, J. Zhu, J. Freeman, P. Dean, A. Valavanis, A. G. Davies, and E. H. Linfield, 'Terahertz quantum cascade lasers with >1 W output powers', *Electron. Lett.*, **50**(4):309–311, 2014.

4

Diffusion

4.1 Introduction

Any substance will attempt to diffuse from an area where it is present in high concentrations to an area of low concentration. For example, obvious though it seems, if the curve $x(z)$ in Fig. 4.1 represented the concentration of water in a trough then the water would fall very rapidly from the region of high concentration to the region of low! Crude though it may seem, this is an example of diffusion. It occurs also for gaseous systems, e.g. smoke gradually disperses in an enclosed room. However, the point of interest in this present work is diffusion of material species in solids.

In the context of semiconductors—and, in particular, semiconductor heterostructures—it is clear that diffusion of *material species* could be important as their very nature derives from discontinuous changes in materials. Figure 4.1 could therefore represent a dopant in either a bulk semiconductor (i.e. a *homojunction*) or an alloy component (e.g. Al at a GaAs/$Ga_{1-x}Al_xAs$ *heterojunction*).

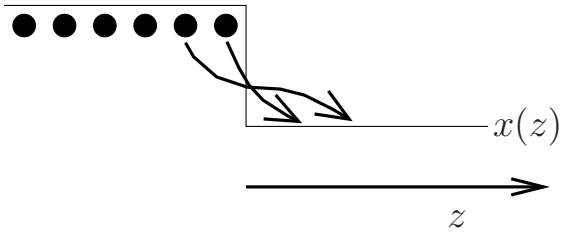


Figure 4.1: Simple illustration of diffusion: the diffusant moves from areas of high concentration to low

Diffusion at such boundaries is a strong possibility, particularly during the elevated temperatures often used during growth, or the lower but more prolonged heating that may occur during normal device operation. Any movement of material (e.g. Al at a GaAs/$Ga_{1-x}Al_xAs$ heterojunction) will 'blur' the interface (i.e. the change from one material to the other will occur via a range of intermediate alloys). Such a process is known as *interface mixing* and is represented schematically in Fig. 4.2. The change in profile of the junction will inevitably alter the electronic properties of the system, which will, in turn,

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

affect the device characteristics and therefore the operating lifetime. The motivation behind modelling diffusion is to understand such time dependency of the properties of quantum well systems, in order to be able to control or prevent it, design a device in which its impact will be minimised, or predict the lifetime of a device. Besides these detrimental effects, controlled diffusion is sometimes initiated deliberately in order to tailor the optical and electronic properties of quantum well systems [1]. For a comprehensive treatment of the use of quantum well interdiffusion for photonics, see the review by Li [2].

Diffusion was first put on a quantitative basis by Fick in 1855 [3], who derived two laws. The first of these stated that the steady-state flux through a plane is proportional to, but in the opposite direction to, the concentration gradient, i.e.

$$\text{flux} = -\mathcal{D}\nabla c \tag{4.1}$$

The second law showed that the time dependency of the concentration is given by:

$$\frac{\partial c}{\partial t} = -\nabla_{\bullet}\text{flux} \tag{4.2}$$

$$\therefore \frac{\partial c}{\partial t} = \nabla_{\bullet}\left(\mathcal{D}\nabla c\right) \tag{4.3}$$

which in one dimension would become:

$$\frac{\partial c}{\partial t} = \frac{\partial}{\partial z}\mathcal{D}\frac{\partial c}{\partial z} \tag{4.4}$$

For a simple introduction to these laws, see Callister [4], p. 66.

Many different particular solutions of Fick's first and second laws can be found, with perhaps the most common one being the error function solution for diffusion, described by a constant $\mathcal{D}$, at the interface between two semi-infinite slabs. Such a model is commonly employed to model diffusion at a semiconductor heterojunction (see [1, 5, 6, 7, 8]).

Crank [9] and Shewmon [10] offer several solutions based upon numerical methods and involving complex initial diffusant profiles and variable diffusion coefficients. However,

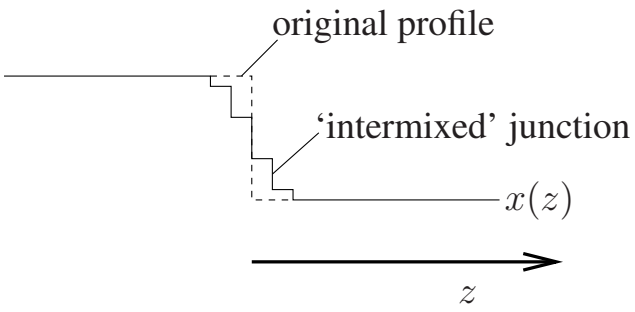


Figure 4.2: An intermixed heterojunction; note that the total amount of diffusant, represented by the area under the $x(z)$ curve, is constant

as with Tuck [11], each situation is considered separately and a new numerical solution is sought. It would be advantageous if a general, probably numerical, method could be developed which was applicable to all of these situations and also new ones.

Such a method would attempt to model diffusion processes, such as those mentioned above, in terms of an average *diffusion coefficient* which may be a constant or a simple function of the material properties. The mechanisms by which the diffusion proceeds, e.g. vacancy-assisted [12], are not of interest. Atomistic models of diffusion *do* exist and are generally centred around a Monte Carlo simulation of the movement of individual atoms within the crystal [13].

For the purposes of this book, solutions are demonstrated within semiconductor heterostructures constructed of an alloy $A_{1-x}B_xC$, where the system can be most generally represented by the z-dependence of the alloy component, i.e. $x = x(z)$. The methods developed are equally applicable to a dopant distribution whose profile may be $c(z)$, as labelled above.

4.2 Theory

As mentioned above, the *most general* one-dimensional diffusion equation for a diffusant distribution represented by $x(z)$ is given by Fick's second law for non-steady-state diffusion [9]:

$$\frac{\partial x}{\partial t} = \frac{\partial}{\partial z}\left(\mathcal{D}\frac{\partial x}{\partial z}\right) \tag{4.5}$$

where t is the time and the diffusion coefficient $\mathcal{D}$ could have temporal t, spatial z and concentration x dependencies, i.e. $\mathcal{D} = \mathcal{D}(x, z, t)$. Given this, the derivative with respect to z operates on both factors, resulting in the following:

$$\frac{\partial x}{\partial t} = \frac{\partial \mathcal{D}}{\partial z}\frac{\partial x}{\partial z} + \mathcal{D}\frac{\partial^2 x}{\partial z^2} \tag{4.6}$$

which is a second-order ($\partial^2/\partial z^2$) nonlinear ($(\partial/\partial z)^2$) differential equation. In any given problem it is likely that two of the following unknowns will be known (!):

- the initial diffusant profile, $x(z, t = 0)$;
- the final diffusant profile, $x(z, t)$;
- the diffusion coefficient, $\mathcal{D} = \mathcal{D}(x, z, t)$.

The problem will be to deduce the third unknown. This could manifest itself in several ways:

a. given the initial diffusant profile and the diffusion coefficient, predict the diffusant profile a certain time into the future;

b. given the initial and final diffusant profiles, calculate the diffusion coefficient;

c. given the final diffusant profile and the diffusion coefficient, calculate the initial diffusant profile.

Knowing the versatility achieved by the *numerical* shooting method solution to Schrödinger's equation as discussed in the previous chapter, clearly a numerical solution would again

be favourable. Learning from the benefits of expanding the derivatives in the Schrödinger equation with finite differences, this would then appear to offer a promising way forward.

Recall the finite- difference approximations to first and second derivatives, i.e.

$$\frac{\partial f}{\partial z} \approx \frac{f(z+\delta z) - f(z-\delta z)}{2\delta z} \tag{4.7}$$

and

$$\frac{\partial^2 f}{\partial z^2} \approx \frac{f(z+\delta z) - 2f(z) + f(z-\delta z)}{(\delta z)^2} \tag{4.8}$$

Then equation (4.6) can be expanded to give:

$$\begin{aligned}\frac{x(z,t+\delta t) - x(z,t)}{\delta t} = &\left[\frac{\mathcal{D}(x,z+\delta z,t) - \mathcal{D}(x,z-\delta z,t)}{2\delta z}\right] \\ &\times \left[\frac{x(z+\delta z,t) - x(z-\delta z,t)}{2\delta z}\right] \\ &+\mathcal{D}(x,z,t)\left[\frac{x(z+\delta z,t) - 2x(z,t) + x(z-\delta z,t)}{(\delta z)^2}\right]\end{aligned} \tag{4.9}$$

Notice that the derivative with respect to time has been written as:

$$\frac{x(z,t+\delta t) - x(z,t)}{\delta t}, \quad \text{rather than} \quad \frac{x(z,t+\delta t) - x(z,t-\delta t)}{2\delta t} \tag{4.10}$$

as would be expected from the expansion in equation (4.7). In fact, in this case where this is the only time derivative, the two are equivalent.

Assuming the most common class of problem, as highlighted above, as point (a), namely that the function $x(z,t)$ is known when $t = 0$, i.e. it is simply the initial profile of the diffusant, and the diffusion coefficient $\mathcal{D}$ is fully prescribed, then it is apparent from equation (4.9) that the concentration x at any point z can be calculated a short time interval δt into the future, provided that the concentration x is known at small spatial steps δz either side of z. This approach to the solution of the differential equation is known as a *numerical simulation*. It is not a mathematical solution, but rather a computational scheme which has been derived to mirror the physical process.

It has already been mentioned that the diffusion coefficient $\mathcal{D}$ could be a function of x, z, and t, with the form of $\mathcal{D}$ being used to define the class of diffusion problem, e.g.

(i). $\mathcal{D} = D_0$, a constant, for simple diffusion problems.

(ii). $\mathcal{D} = D(x)$, a function of the concentration as encountered in nonlinear diffusion problems [14]. Note that as $x = x(z)$, then $\mathcal{D}$ is intrinsically a function of position too.

(iii). $\mathcal{D} = D(z)$, a function of position only, as could occur in ion implantation problems [15]. Here, the diffusion coefficient could be linearly dependent on the concentration of vacancies for example, where the latter itself is depth-dependent.

(iv). $\mathcal{D} = D(t)$, a function of time, as could occur during the annealing of radiation damage. For example, ion implantation can produce vacancies which aid diffusion [16, 17]. During an anneal, the vacancy concentration decreases as the lattice is repaired, which in turn alters the diffusion coefficient.

4.3 Boundary conditions

Thus, given the initial diffusant profile and a fully prescribed diffusion coefficient, everything is in place for predicting the profile of the diffusant at any time in the future, except for the conditions at the ends of the system. These cannot be calculated with the iterative equation (equation (4.9)), as it requires points that lie outside the z-domain.

For diffusion from an infinite source, it may be appropriate to fix the diffusant concentration x at the two end points, e.g. $x(z = 0, t) = x(z = 0, t = 0)$. Alternatively, the concentrations x at the limits of the z-domain could be set equal to the adjacent points, which can be deduced from equation (4.9). Physically, this defines the semiconductor structure as a closed system, with the total amount of diffusant remaining the same. It is these latter 'closed system' boundary conditions that will be employed exclusively in the following examples.

4.4 Convergence tests

Figure 4.3 shows the result of allowing the diffusant profile in Fig. 4.1 to evolve to equilibrium, using the closed-system boundary conditions as described above. The 'closed' nature of the system can clearly be seen—the total amount of diffusant remains the same, and ultimately, as would be expected for the 'water step', the concentration reaches a constant value. In the case of water, this could be looked upon as minimising the potential energy.

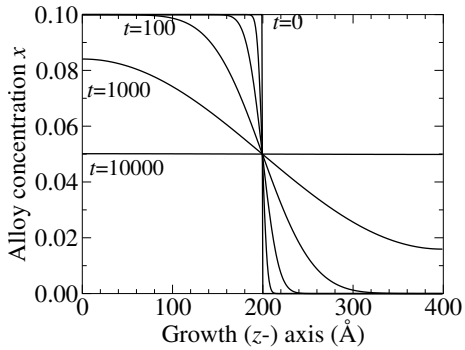


Figure 4.3: Time evolution of adjacent finite slabs of $Ga_{0.9}Al_{0.1}As$/GaAs with a constant diffusion coefficient of $\mathcal{D} = D_0 = 10\,\text{Å}^2\,\text{s}^{-1}$

If the diffusion process can be described by a constant diffusion coefficient, $\mathcal{D} = D_0$, then the general diffusion equation (equation (4.5)) and its equivalent computational form (equation (4.9)) simplify to the following:

$$\frac{\partial x}{\partial t} = D_0 \frac{\partial^2 x}{\partial z^2} \tag{4.11}$$

which, as mentioned above, has error function solutions for the case of diffusion at the interface ($z = 0$) of a semi-infinite slab of concentration x_0 [9], i.e.

$$x(z) = \frac{x_0}{2}\text{erfc}\frac{z}{2\sqrt{D_0 t}} \tag{4.12}$$

where erfc is the complementary error function (see [18], p. 295). This technique can be imposed on multiple heterojunctions by linearly superposing solutions [5]. Figure 4.4 compares just such an error function solution[1] with the numerical solution for a 200 Å single GaAs quantum well surrounded by 200 Å $Ga_{0.9}Al_{0.1}As$ barriers after 100 s of diffusion described by a constant coefficient $D_0 = 10\,\text{Å}^2\,\text{s}^{-1}$. Clearly, the numerical method advocated here exactly reproduces the analytical solution.

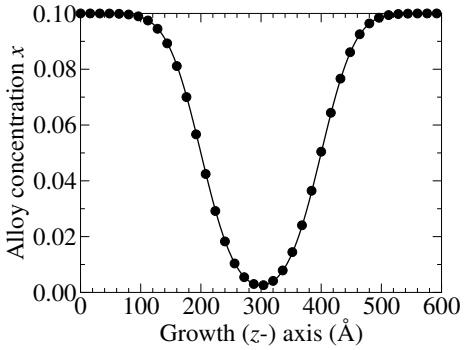


Figure 4.4: Comparison between the numerical solution (solid symbols) and the standard error function solution (continuous line) for the concentration profile $x(z)$ of a 200 Å $Ga_{0.9}Al_{0.1}As$/200 Å GaAs /200 Å $Ga_{0.9}Al_{0.1}As$ single quantum well, after 100 s of diffusion described by a constant diffusion coefficient $D_0 = 10\,\text{Å}^2\,\text{s}^{-1}$

As with any numerical method, convergence tests must be performed. In the present case, these are necessary in order to determine that the results are independent of the intervals δz and δt. For the characteristic dimensions of interest in the following problems, namely nanostructures ($\sim$100–1000 Å) and macroscopic annealing times ($\sim$100–1000 s), $\delta z = 1$ Å and $\delta t = 0.01$ s were found to be satisfactory.

From a computational viewpoint, the technique is useful since, in general, it is anticipated that the numerical complexity scales linearly with the physical complexity. In actuality, it has been found that, for the series of problems discussed below, the number of spatial points required for the calculation is independent of the variability in the initial concentration profile. Larger microstructures (i.e. superlattices with more and more periods) do require more points (i.e. δz remains constant), but the computational effort required is proportional to the number of points. For a given structure, but with an increasingly complex dependence of the diffusion coefficient (e.g. $\mathcal{D}$ having a stronger functional dependence on the concentration

[1] Thanks are due to T. Stirner for providing the error function data.

x), smaller time intervals δt are required. Again, however, halving the time interval merely doubles the necessary computational effort.

Having therefore established a numerical method for simulating diffusion and after validating it by comparison with the accepted analytical form for a special case, it is now an appropriate time to demonstrate the versatility of the solution by visiting, in turn, examples (i)–(iv) listed in Section 4.2.

4.5 Numerical stability

The finite-difference approximation in equation (4.9) is known as a forward-time central-space formulation, since the value of $x(z,t)$ is projected forward in time from a known value, and the spatial derivative is 'pivoted' around a given point in space. Forward-time central-space schemes are relatively simple to implement computationally, but they are not unconditionally stable. It can be shown that a stable solution is only achieved if the time and space intervals satisfy the following condition [19]:

$$\frac{D_0 \delta t}{\delta z^2} \leq \frac{1}{2} \tag{4.13}$$

If this condition is not satisfied, then the algorithm diverges away from the true solution at each time-step, and $x(z,t)$ reaches a large, physically unrealistic value. Therefore, any increase in the spatial resolution of the simulation must also be accompanied by a *squared* increase in the time resolution, in order to guarantee stability. As such, this can lead to very long simulation times for systems with narrow features.

A more robust solution to the diffusion equation can be found using unconditionally stable (although more complex) techniques such as the Crank–Nicolson method [19]. For the purposes of this chapter, however, it suffices to check the stability criterion above for each simulated structure.

4.6 Constant diffusion coefficients

The case of a constant diffusion coefficient has already been touched upon in the previous section, as its analytical solutions represent an important test for the numerical solution forwarded here. Such a numerical solution can be used to illustrate two specific points for this situation.

In particular, the error function solutions given by equation (4.12) imply *universality*, in that they are dependent upon the value of the product $D_0 t$ only, and not the individual values themselves. Therefore, a system with a high diffusion coefficient can give the same diffused profile after a short time as a system with a low diffusion coefficient that is allowed to diffuse for longer periods. Indeed, equation (4.12) is often written in terms of a *diffusion length* property [20], $L_\mathrm{d} = 2\sqrt{D_0 t}$, such that:

$$x(z) = \frac{x_0}{2} \mathrm{erfc} \frac{z}{L_\mathrm{d}} \tag{4.14}$$

The continuous lines in Fig. 4.5 display the results of diffusing a single quantum well for 100 (bottom) and 1000 s (top curve) with $D_0 = 10\,\text{Å}^2\,\mathrm{s}^{-1}$. In comparison with this,

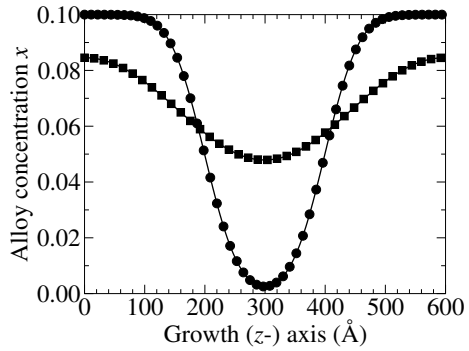


Figure 4.5: Diffusion profiles for a 200 Å $Ga_{0.9}Al_{0.1}As$/200 Å GaAs /200 Å $Ga_{0.9}Al_{0.1}As$ single quantum well, with constant diffusion coefficient D_0, and $D_0 t = 1000\,\text{Å}^2$ (continuous line plus solid circles) and $D_0 t = 10\,000\,\text{Å}^2$ (continuous line plus solid squares)

the solid symbols are the results of diffusing the same initial system for just $10\,\text{s}$ but with $D_0 = 100$ (circles) and $1000\,\text{Å}^2\,\text{s}^{-1}$ (squares). Clearly, the numerical solution reproduces the universality as well—a point that is not at all obvious from the numerical form of the diffusion equation (4.9).

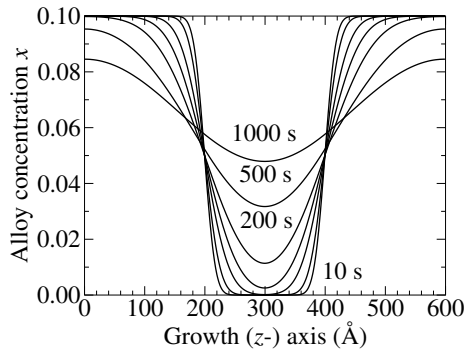


Figure 4.6: Diffusion profiles for the single quantum well of Fig. 4.5 with constant diffusion coefficient $D_0 = 10\,\text{Å}^2\,\text{s}^{-1}$, and diffusing times $t = 10$, 20, 50, 100, 200, 500 and $1000\,\text{s}$

Figure 4.6 displays a succession of diffusant profiles for the same single quantum well as before, again with $D_0 = 10\,\text{Å}^2\,\text{s}^{-1}$, where the evolution of the profile at longer diffusing times is clear. Therefore, given a semiconductor heterostructure and an annealing time, it is apparent that the diffused profile can be predicted, *provided* that the diffusion coefficient characterising the process is known from other experiments. The direct determination of the diffused profile on such short length scales is difficult (although not impossible) from

traditional techniques used for bulk analysis, such as *secondary ion mass spectroscopy* (SIMS) (see [21], Section 2.4). However, alternative spectroscopic techniques can be used to infer the extent of diffusion by its effect on other observables. In particular, as mentioned in Chapter 3, photoluminescence and photoluminescence excitation can be used to measure the excitation energy of electrons across the band gap. As the electron and hole confinement energies, E_{e} and E_{h}, are dependent upon the band-edge profiles of the conduction- and valence-band edges, respectively, the excitation energy ($E = E_{\mathrm{g}} + E_{\mathrm{e}} + E_{\mathrm{h}}$, ignoring the exciton binding energy), is then also dependent upon the amount of diffusion.

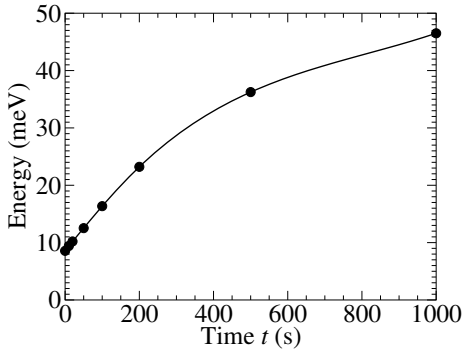


Figure 4.7: The corresponding ground state electron energies for the diffused profiles given in Fig. 4.6

The curve in Fig. 4.7 illustrates the change in the electron energy corresponding to the diffusion process of Fig. 4.6. The magnitude of the variation is substantial and easily detectable by experiment [6, 7, 8, 22, 23, 24]. For a theoretical interpretation of such experiments, *including* the change in exciton binding energy, see [5].

4.7 Concentration-dependent diffusion coefficient

There are many examples for which the diffusion coefficient of a species is a function of the concentration of that species itself [11, 25, 26].

The work by Tuck [11] is perhaps the most interesting as it makes a detailed comparison with experiment. Tuck considered the diffusant of a dopant, Zn in bulk GaAs, a well-studied system in which the 'substitutional-interstitial' mechanism is thought to dominate [27]. In this system the dopant (Zn) sits at two different positions within the crystal lattice, (i.e., as the name implies, both substitutionally and at interstitial sites). Atoms on the substitutional sites, which are in the majority, diffuse very slowly; however, the minority of interstitial atoms diffuse so readily that they dominate the evolution of the dopant profile with time.

The diffusion coefficient describing the process is dependent on concentration (c), and is given mathematically as:

$$\mathcal{D} = kc^n \tag{4.15}$$

hence the diffusion equation becomes:

$$\frac{\partial c}{\partial t} = \frac{\partial}{\partial z}\left(kc^n \frac{\partial c}{\partial z}\right) \tag{4.16}$$

The constant k was deduced from experiment, and the index n, for this system, was found to be 2. Tuck considered several situations and implemented solutions which agreed very well with experiment. In this work, the numerical simulation can be implemented directly.

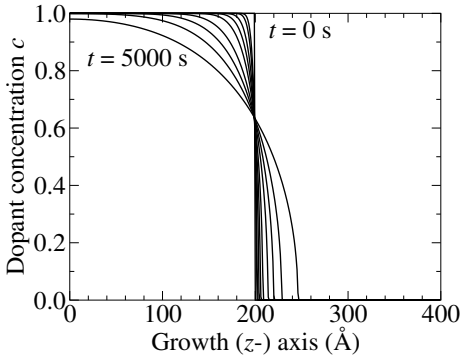


Figure 4.8: Concentration-dependent ($\mathcal{D} = kc^2$) diffusion

Figure 4.8 presents examples of the concentration profiles that such a concentration-dependent diffusion coefficient yields. The initial concentration c was just taken to be unity and k was taken as $1\,\text{Å}^2\,\text{s}^{-1}$. Note that, unlike the case of the constant diffusion coefficient, the curves do not all pass through the midpoint of the initial profile height, as shown in Fig. 4.3. Most importantly, the curves are of the same form as those given in Tuck [11] and hence agree with experiment.

4.8 Depth-dependent diffusion coefficient

As mentioned above, diffusion is becoming widely used as a post-growth method for fine-tuning the structural, and hence electronic and optical, properties of semiconductor heterostructures. Ion implantation (see Kelly [21], Section 3.4, for an introduction) is one method of stimulating diffusion; it is controllable, reproducible and its spatial resolution can be used to pattern a semiconductor wafer. Ion implantation can also be used as a way of introducing new species into a crystal, e.g. optically active rare earth ions into a host semiconductor [28]. However, the focus of interest here is on ion implantation as a means of enhancing the diffusion of the material species already present [17, 22, 24]. The implantation itself produces lattice damage, displacing atoms from their equilibrium positions and thus creating interstitial atoms and vacancies. As in the previous example, the interstitial atoms can diffuse very readily, and in addition the presence of vacancies provides a route for the diffusion of substitutional species.

The diffusion coefficient describes the average speed by which atoms diffuse, and clearly the more vacancies and interstitials in a given region of material, the faster diffusion will proceed. Hence, in order to attempt to quantify the diffusion coefficient, it is necessary to have a knowledge of the amount of lattice damage that an implantation produces; this can be calculated with a Monte Carlo simulation of the atoms impinging on the crystal planes. Such computer simulations, known as TRIM codes, are well understood and documented (see, for example, [29]). An example of the output from such a simulation is shown as the solid line in Fig. 4.9.

The effect of 'vacancy-assisted' diffusion from such a profile has been calculated for a CdTe/$Cd_{1-x}Mn_xTe$ multiple quantum well [15, 30]. In this work, the depth dependence will be exaggerated merely to produce a more interesting variation in the diffusion. The depth dependence of the vacancy concentration could actually be described *empirically* by a Gaussian distribution of the form:

$$\rho^{\text{vacancy}} = \rho_0^{\text{vacancy}} \exp\left[-\frac{(z-z_0)^2}{2\sigma^2}\right] \tag{4.17}$$

where z_0 is the depth at which the maximum vacancy concentration occurs, and the standard deviation σ describes the width of the distribution. Figure 4.9 displays such a Gaussian fit to the TRIM data, described by the parameters $z_0 \sim 1800\,\text{Å}$ and $\sigma = 1700$ Å, together with a modified distribution for the purpose of this example, given by $\sigma = 600$ Å.

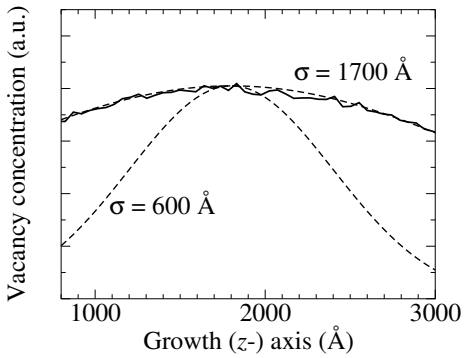


Figure 4.9: TRIM code data (solid bold line) for $200\,\text{keV}$ Ar^{2+} ions impinging on a [001] CdTe surface. A Gaussian fit to the data is shown as a dashed line, together with another Gaussian with exaggerated depth dependence

In real systems where intensive investigations for a particular semiconductor multilayer with a particular ion implantation dosage are carried out, it would be necessary to relate the diffusion coefficient to the absolute vacancy concentration; however, in these present demonstrator examples of the numerical solution to the diffusion equation, it suffices to let $\mathcal{D}$ be proportional to ρ^{vacancy}, and choose the constant of proportionality to be $10\,\text{Å}^2\,\text{s}^{-1}$, i.e.

$$\mathcal{D} = 10 \exp\left[-\frac{(z-z_0)^2}{2\sigma^2}\right] \text{Å}^2\,\text{s}^{-1} \tag{4.18}$$

Given this, Fig. 4.10 displays both the initial and final alloy concentration profile after $200\,\mathrm{s}$ of diffusion described by equation (4.18) for a generic $150\,\text{Å}$ AC/$50\,\text{Å}$ $\mathrm{A}_{1-x}\mathrm{B}_x\mathrm{C}$ superlattice/multiple quantum well. As expected, the central wells have diffused considerably more than those near the edges where the diffusion coefficient is lower. The exaggerated z-dependence of $\mathcal{D}$ has fulfilled its goal in this illustration of producing a much clearer depth dependence than previously published for a realistic system (see, for example, [30]).

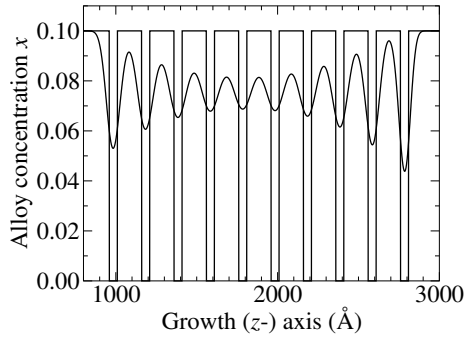


Figure 4.10: Alloy concentration profile x after $200\,\mathrm{s}$ of diffusion described by the depth dependent coefficient of equation (4.18)

Photoluminescence measurements on diffused systems such as those shown in Fig. 4.10 would exhibit a broadened emission line, as the photogenerated carriers in the central wells would have a different energy than those in the outer wells. This phenomenon of line broadening in a diffused superlattice has been observed by Elman *et al*. [22] and subsequently modelled theoretically [30].

Alternatively, the structure, including the depth dependence of the diffusion, can be mapped directly by using *double-crystal X-ray diffraction* [31], a technique which has shown itself to be very valuable in monitoring the progress of interface intermixing [17].

4.9 Time-dependent diffusion coefficient

Vacancy-enhanced diffusion will continue for as long as the vacancies are present, but one way to control their lifetime and 'freeze' the diffusant profile is to anneal out the radiation damage. The annealing process thermally activates the interstitials back into vacancies and so restores order to the crystal. This subsequent annealing process *could* be described by a time-dependent diffusion coefficient, perhaps of the form:

$$\mathcal{D} = \mathcal{D}(z)\exp\left(-\frac{t}{\tau}\right) \tag{4.19}$$

thus presenting the opportunity to complete the examples of the functional dependencies listed in Section 4.2. $\mathcal{D}(z)$ is the initial (time-independent) depth-dependent diffusion

coefficient due to the vacancy distribution. Using the form in equation (4.18), the diffusion coefficient would then be fully specified by:

$$\mathcal{D} = 10 \exp\left[-\frac{(z-z_0)^2}{2\sigma^2}\right] \exp\left(-\frac{t}{\tau}\right) \text{Å}^2\,\text{s}^{-1} \tag{4.20}$$

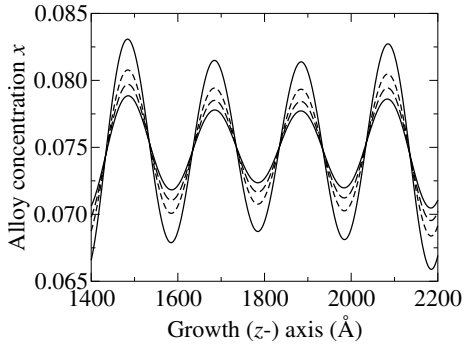


Figure 4.11: Annealing out radiation damage—an example of time-dependent diffusion, with annealing times of 0, 50, 100 and $200\,\text{s}$ giving the concentration profiles of *decreasing* amplitude

Making use of the ion-implantation-enhanced diffusion profile of Fig. 4.10 as a starting point, Fig. 4.11 shows the results of simulating annealing out the lattice damage with a time-dependent diffusion coefficient of the form shown in equation (4.20). In this case, the decay time τ of the vacancy concentration was taken to be $100\,\text{s}$, and so the curves represent the points at which the vacancy concentration is a fraction 1, $\text{e}^{-\frac{1}{2}}$, e^{-1}, and e^{-2} of its original value. Clearly, the curves are converging to a point, which represents the region at which the ion-implantation-enhanced diffusion has been frozen.

4.10 δ-doped quantum wells

In the previous sections, examples have been given of simulating diffusion for all of the various forms of diffusion coefficient that can exist. It is clear that the computation method can be extended to include combinations of all three dependencies, and indeed in the final example the diffusion coefficient had both depth and time dependency. It now serves a purpose to follow through an example of a diffusion problem that is of direct relevance to semiconductor heterostructures, and to calculate the subsequent effects on an observable, in this case, the quantum-confinement energies.

Contemporary epitaxial growth techniques, such as molecular beam epitaxy and chemical beam epitaxy, allow for the possibility of growing very thin layers of semiconductor material. Another possibility is the potential for these techniques to lay down very thin layers of dopant atoms. Figure 4.12(a) represents the bulk doping of a layer, as used in HEMTs and

the majority of semiconductor heterostructure devices. Figure 4.12(b) represents the dopant profile for a single δ-layer in a quantum well, although of course it could be in a barrier or at an interface. Such thin layers of dopant are called *delta-layers*, as the dopant profile resembles the Dirac δ-function, and the whole process is known as δ-doping; for a comprehensive treatise of this technique, see Schubert [32].

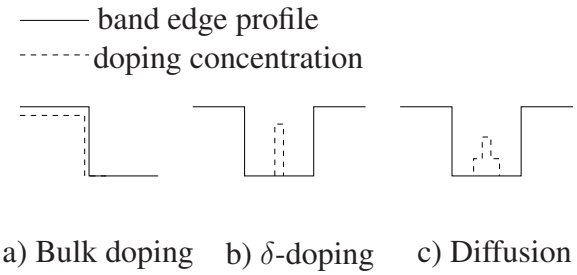


Figure 4.12: Doping and δ-doping

One of the main problems affecting δ-doped layers is diffusion of the dopant which can often occur with the high temperatures employed during growth. A δ-layer designed to extend over just one monolayer can in fact spread over several layers. This would affect the electronic and optical properties of the device, in particular the electron or hole confinement energies, and scattering from the distribution of ionised impurities would be different. In this section, the focus will be on the first of these points, as the techniques to model such a system have already been covered in this work.

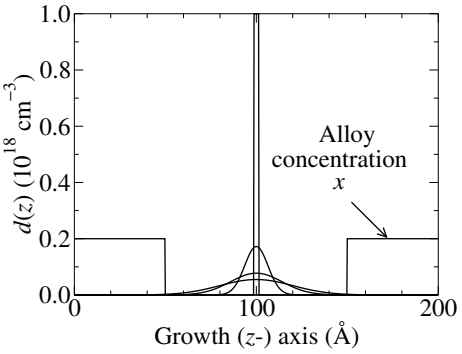


Figure 4.13: Dopant concentration profile, $d(z)$, after 0, 20, 100 and 200 s (in order of decreasing maximum height) of diffusion described by a constant diffusion coefficient of $1\,\text{Å}^2\,\text{s}^{-1}$, together with the alloy concentration x of the $Ga_{1-x}Al_xAs/GaAs/Ga_{1-x}Al_xAs$ single quantum well

Figure 4.13 displays the dopant profile of a single-monolayer-thick δ-layer when subject to diffusion described by a constant coefficient $D_0 = 1\,\text{Å}^2\,\text{s}^{-1}$. Clearly, the total amount of dopant remains constant, and hence as the profile broadens, the height of the peak decreases. If the dopant were p-type, e.g. beryllium in GaAs, then at elevated temperatures the acceptors would be ionised and donate holes which would become localised in the quantum well. By using the methods outlined in Chapter 3, the Coulomb potential due to this negative acceptor/positive hole distribution can be calculated and incorporated in the self-consistent solution of Schrödinger's and Poisson's equations.

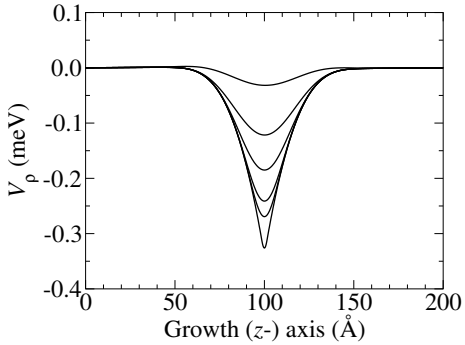


Figure 4.14: The self-consistent potential V_ρ due to the acceptor/hole distribution in the δ-doped quantum well of Fig. 4.13, for diffusion times, in order of decreasing depth, of 0, 10, 20, 50, 100 and $200\,\text{s}$

Assuming that all of the holes occupy the lowest heavy-hole subband in the quantum well, then Fig. 4.14 shows the self-consistent potential V_ρ due to the charge distribution for various diffusion times. Clearly, as the δ-doped layer broadens, the depth of the potential decreases. Interestingly, the potential seems to tend towards zero; this is because the diffused acceptor distribution begins to resemble the self-consistent hole wave functions, thus leading to an almost complete cancellation of the positive and negative potentials.

Figure 4.15 displays the self-consistent heavy-hole energy as a function of the diffusion time. At this relatively low doping density, the effect of diffusion on the heavy-hole energy is small; nonetheless, this example serves the purpose of demonstrating the applicability of the diffusion equation solution to systems of interest for devices, while the inclusion of the self-consistent Schrödinger and Poisson solution shows the power of combining such techniques together. Note that, as a constant diffusion coefficient was employed, all of the curves are universal in $D_0 t$.

4.11 Extension to higher dimensions

The one-dimensional diffusion equation can be generalised to higher dimensions [9], i.e.

$$\frac{\partial x}{\partial t} = \nabla_\bullet \left(\mathcal{D} \nabla x\right) \tag{4.21}$$

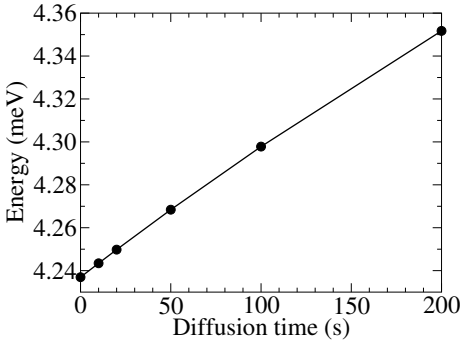


Figure 4.15: The self-consistent heavy-hole energy as a function of the diffusion time

This can again be expanded in terms of finite differences to give an iterative equation in analogy with equation (4.9). This time, however, the concentration at a point for a time interval δt into the future is dependent on the diffusion coefficient $\mathcal{D}$ and the concentration x, which may both be functions of all of the spatial coordinates. Therefore, an iteration needs to be performed for each spatial dimension. Following this, an important check must be made that the resulting new concentration is independent of the order in which the iterations were performed.

Exercises

(1). The analytical solution to diffusion from a semi-infinite slab of initial concentration x_0 across a boundary at $z = 0$ into a region of initially zero concentration with a constant diffusion coefficient D_0 is a particular kind of error function. Write down the analytical solution for describing how x varies over time t along the z-axis. Show how the solutions can be expressed in terms of a diffusion length. Explain why this approach may be particularly useful for the analysis and modelling of experimentally grown structures, when the details of the diffusion processes during growth are poorly understood.

(2). Show, by substituting in, that the error function solution in the previous question satisfies Fick's second law with a constant diffusion coefficient D_0 , i.e.

$$\frac{\partial x}{\partial t} = D_0 \frac{\partial^2 x}{\partial z^2}$$

(3). Find the first few terms in the Taylor series expansion of the error function $\mathrm{erf}(u)$, where u is the independent variable. Hence deduce the equivalent for the complementary error function $\mathrm{erfc}(u)$. Show that near the boundary between the semi-infinite slabs of the previous questions, for small diffusion cofficients and short times, the solution to the previous question can be approximated as:

$$x(z,t) = \frac{x_0}{2}\left(1 - \frac{z}{\sqrt{D_0 t \pi}}\right)$$

(4). Show that solutions of Fick's second law of diffusion are linear and can be superposed, i.e. the solution of a linear combination of initial distributions $x(z, t = 0)$ can be constructed from a linear combination of the solutions of individual initial distributions.

(5). Use the principle of superposition discussed in the previous question to write a computer code to produce the complementary error function solution of diffusion in the finite quantum well of Fig. 4.4.

(6). Use the `qwwad_diffuse` and `qwwad_ef_generic` programs to investigate interdiffusion in a system consisting of $60\,\text{Å}$ and $40\,\text{Å}$ GaAs quantum wells separated by a $30\,\text{Å}$ $Ga_{0.8}Al_{0.2}As$ barrier. Assume a constant $D_0 = 10\,\text{Å}^2\,\text{s}^{-1}$ diffusion coefficient. Determine the effect on the energy separation between states, and the extent of the wave functions, as diffusion time increases. Explain this behaviour qualitatively, and discuss the implications of alloy diffusion for devices such as quantum well lasers whose emission wavelength depends on the precise separation between subbands.

References

[1] E. H. Li, 'Interdiffusion as a means of fabricating parabolic quantum wells for the enhancement of the nonlinear third-order susceptibility by triple resonance', *Appl. Phys. Lett.*, **69**:460–462, 1996.

[2] E. H. Li, Ed., *Selected Papers on Quantum Well Intermixing for Photonics*, SPIE Optical Engineering Press, Bellingham, WA, 1998.

[3] A. Fick, 'Generalized Fick law for anomalous diffusion', *Ann. Phys.*, **170**:59, 1855.

[4] W. D. Callister, Jr., *Materials Science and Engineering*, John Wiley & Sons, Inc., New York, 1985.

[5] P. Harrison, W. E. Hagston, and T. Stirner, 'Excitons in diffused quantum wells', *Phys. Rev. B*, **47**:16404–16409, 1993.

[6] P. Kossacki, N. T. Khoi, J. A. Gaj, G. Karczewski, T. Wojtowicz, E. Janik, A. Zakrzewski, M. Kutrowski, and J. Kossut, 'Rapid thermal processing of semimagnetic superstructures studied by magnetoreflectivity', *Superlatt. Microstruct.*, **16**(1):63–66, 1994.

[7] M. T. Furtado and M. S. S. Loural, 'Direct evaluation of interdiffusion coefficients in quantum well heterostructures using photoluminescence', *Superlatt. Microstruct.*, **14**(1):21, 1993.

[8] D. Tönnies, G. Bacher, A. Forchel, A. Waag, and G. Landwehr, 'Photoluminescence study of strong interdiffusion in CdTe/CdMnTe quantum wells induced by rapid thermal annealing', *Appl. Phys. Lett.*, **64**:766, 1994.

[9] J. Crank, *The Mathematics of Diffusion*, Oxford University Press, Oxford, 1956.

[10] P. Shewmon, *Diffusion in Solids*, McGraw-Hill, New York, 1963.

[11] B. Tuck, 'Some explicit solutions to the non-linear diffusion equation', *J. Phys. D: Appl. Phys.*, **9**:123, 1976.

[12] D. Shaw, 'Diffusion mechanisms in II–VI materials', *J. Crystal Growth*, **86**(1–4):778–796, 1988.

[13] K. Binder, 'Atomistic modeling of materials properties by Monte Carlo simulation', *Adv. Mater.*, **4**(9):540–547, 1992.

[14] P. Harrison, 'Differentiating between constant and concentration-dependent diffusion coefficients via the optical spectroscopy of excitons in quantum wells', *Semicond. Sci. Technol.*, **11**:1022–1025, 1996.

[15] J. M. Fatah, I. Karla, P. Harrison, T. Stirner, W. E. Hagston, and J. H. C. Hogg, 'Defect induced diffusion mechanisms in ion implanted quantum well structures', in *Proceedings of the 22nd International Conference on the Physics of Semiconductors*, Singapore, 1994, pp. 2275–2278, World Scientific.

[16] I. Karla, D. Shaw, W. E. Hagston, J. H. C. Hogg, S. Chalk, J. E. Nicholls, and C. Peili, 'Measurement of interdiffusion in II–VI quantum-well structures using optical methods', *J. Appl. Phys.*, **79**(4):1895–1897, 1996.

[17] I. Karla, J. H. C. Hogg, W. E. Hagston, J. Fatah, and D. Shaw, 'Monitoring of intermixing and interdiffusion by x-ray diffraction of ion-implanted quantum-well structures', *J. Appl. Phys.*, **79**(4):1898–1902, 1996.

[18] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover Publications, New York, 1965.

[19] J. D. Hoffman, *Numerical Methods for Engineers and Scientists*, Marcel Dekker, New York, 2nd edition, 2001.

[20] E. Li, B. L. Weiss, and K.-S. Chan, 'Eigenstates and absorption spectra of interdiffused AlGaAs–GaAs multiple-quantum-well structures', *IEEE J. Quantum Electron.*, **32**(8):1399–1416, 1996.

[21] M. J. Kelly, *Low Dimensional Semiconductors: Materials, Physics, Technology, Devices*, Clarendon Press, Oxford, 1995.

[22] B. Elman, E. S. Koteles, P. Melman, and C. A. Armiento, 'GaAs/AlGaAs quantum-well intermixing using shallow ion implantation and rapid thermal annealing', *J. Appl. Phys.*, **66**(5):2104–2107, 1989.

[23] M. K. Chai, S. F. Wee, K. P. Homewood, W. P. Gillin, T. Cloitre, and R. L. Aulombard, 'An optical study of interdiffusion in ZnSe/ZnCdSe', *Appl. Phys. Lett.*, **69**:1579, 1996.

[24] I. V. Bradley, W. P. Gillin, K. P. Homewood, and R. P. Webb, 'The effects of ion implantation on the interdiffusion coefficients in $In_{1-x}Ga_xAs$/GaAs quantum well structures', *J. Appl. Phys.*, **73**:1686, 1993.

[25] M. A. Litovskii and R. S. Malkovich, 'Method of determining the diffusion profile for a concentration dependent diffusion coefficient', *phys. stat. sol. (a)*, **36**(2):K145–K148, 1976.

[26] A. D. Pelton and T. H. Etsell, 'Analytical solution of Fick's second law when the diffusion coefficient varies directly as concentration', *Acta Metallurgica*, **20**:1269, 1972.

[27] F. C. Frank and D. Turnbull, 'Mechanisms of diffusion of copper in germanium', *Phys. Rev.*, **104**:617, 1956.

[28] T. Taskin, S. Gardelis, J. Evans, B. Hamilton, and A. Peaker, 'Sharp 1.54 μm luminescence from porous erbium implanted silicon', *Electron. Lett.*, **31**(24):2132–2133, 1995.

[29] J. F. Ziegler, J. P. Bierrack, and U. Littmark, *The Stopping and Range of Ions in Matter*, vol. 1, Pergamon Press, New York, 1985.

[30] P. Harrison, 'Numerical solution to the general one-dimensional diffusion equation in semiconductor heterostructures', *phys. stat. sol. (b)*, **197**:81–90, 1996.

[31] J. M. Fatah, P. Harrison, T. Stirner, J. H. C. Hogg, and W. E. Hagston, 'Double crystal X-ray diffraction simulation of diffusion in semiconductor microstructures', *J. Appl. Phys.*, **83**:4037–4041, 1998.

[32] E. F. Schubert, Ed., *Delta-Doping of Semiconductors*, Cambridge University Press, Cambridge, 1996.

5

Impurities

5.1 Donors and acceptors in bulk material

The ability to introduce impurities directly into the lattice was the most important technological advance in the development of semiconductors as electronic materials. This allows their electronic properties to be tailored to suit the engineer's needs. Atoms that are introduced with an additional electron more than that required to form the chemical bonds with neighbouring atoms are easily ionised, thus donating electrons to the crystal. Such atoms are known as donors. Alternatively, impurity atoms can be incorporated into the lattice, which are an electron short of that needed to form the chemical bonds. These atoms are able to accept electron from nearby bonds, thus generating an empty state in the valence band. Atoms of this type are known as acceptors. This *hole* is then mobile within the lattice and can contribute to the conductivity. Doping different regions of the same lattice with donors and acceptors leads to a p-n junction, which was the basis of the first transistor—the bipolar.

The interest here is not in the transport properties of doped semiconductors and their exploitation as the electronic materials of commercial devices such as diodes and transistors, as this has been covered very successfully by many other authors (see, for example, Sze [1]). The interest here is in the properties of impurity states within semiconductor heterostructures. These states have a bearing on the electronic and optical properties of these materials, and again it is the evaluation of the fundamentals governing their behaviour that is the primary concern.

Figure 5.1 shows a schematic representation of a neutral (occupied) donor in a bulk semiconductor. The regular lattice represents a $\{100\}$ plane. An example of a donor in silicon is the group V element phosphorus. In a compound semiconductor such as gallium arsenide or cadmium telluride, the dopant would occupy a cation site. Thus the atoms of Fig. 5.1 would be gallium or cadmium, respectively. In GaAs, a typical donor is silicon, where the latter is forced to sit on the Ga site by growth under an As overpressure. The group IV element Si has one electron more than the group III element Ga and it donates this to the lattice. Note that a Si atom occupying an As site would be an acceptor.

It can be seen from the figure that the unionised donor resembles the electron–proton pair constituting a hydrogen atom; the physics of this is well understood and was initially described with a semi-classical model by Bohr [2] and reinforced later with full quantum mechanical descriptions [3, 4].

The binding energy of this neutral donor–electron pair is given by:

$$E_{\mathrm{D}^0} = -\frac{m_{\mathrm{e}}^* e^4}{32\pi^2\hbar^2\epsilon_{\mathrm{r}}^2\epsilon_0^2} \tag{5.1}$$

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

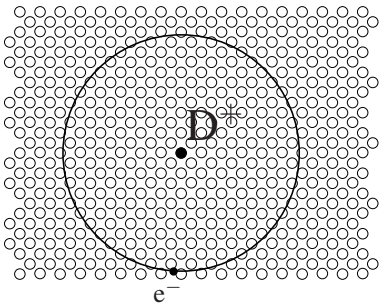


Figure 5.1: Schematic representation of a donor in a bulk semiconductor

and the Bohr radius follows as:

$$\lambda = \frac{4\pi\epsilon_{\mathrm{r}}\epsilon_0\hbar^2}{m_{\mathrm{e}}^* e^2} \tag{5.2}$$

In this case the mass of the donor atom is much larger than the electron mass and hence the simpler 'infinite nucleus mass' variant can be employed. Later, in the chapter dealing with excitons (Chapter 6), this hydrogenic model will again be employed; however, on that occasion the hole mass is of the same order of magnitude as the electron mass and therefore cannot be ignored. Note that the only material parameters needed are the electron effective mass and the permittivity of the host material, and thus the neutral donor binding energy should be independent of the impurity atom. Experimental data reproduced by Sze [5] substantiate this result, for the majority of cases in GaAs.

Taking typical values for GaAs, e.g. the Γ valley electron effective mass $m_{\mathrm{e}}^* = 0.067m_0$ and the static dielectric constant $\epsilon_{\mathrm{r}} = 13.18$ [6], then:

$$E_{\mathrm{D}^0} = -5.3\,\mathrm{meV} \quad \text{and} \quad \lambda = 103\,\text{Å}$$

which is very close to the measured value of $5.8\,\mathrm{meV}$ [5]. Correspondingly, for CdTe, $m_{\mathrm{e}}^* = 0.096m_0$ and $\epsilon_{\mathrm{r}} = 10.6$, and therefore:

$$E_{\mathrm{D}^0} = -11.7\,\mathrm{meV} \quad \text{and} \quad \lambda = 58\,\text{Å}$$

The Bohr radii produced by these simple calculations represent an important result. The lattice constant of GaAs is 5.65 Å, and hence 103 Å represents around 18 unit cells along any radius from the donor to the electron orbit. Or, alternatively, each face-centred cubic unit cell contains four lattice points, or eight atoms, and has a volume of $A_0^3 = (5.65\,\text{Å})^3$, and hence the volume occupied by one atom is $A_0^3/8$. Thus the number of atoms within the spherical electron orbit of radius λ is given by:

$$\text{number of atoms} = \frac{4\pi\lambda^3/3}{A_0^3/8} \tag{5.3}$$

which for GaAs gives 203 000 atoms. This seems like quite a large number and hence provides justification for using the bulk value of the permittivity ϵ_{r}. For systems with a much

smaller Bohr radius, which might contain substantially fewer atoms, it would be apparent that the electromagnetic properties of the crystal are quite different from the bulk, and care must be taken in choosing the value for the permittivity. Harrison *et al*. [7] have shown, by careful comparison of exciton binding calculations with detailed experimental work, that the permittivity required to produce agreement lies between the static, ϵ_s, and infinite, ϵ_∞, frequency values. The binding energies themselves are small relative to the band gap and, obviously, negative, which implies that they lie just below the conduction-band edge.

At low temperatures, there are few lattice vibrations (phonons) and hence the electrons remain bound to the donors. However, as the temperature is increased, the number of phonons within the lattice increases and the donors can become ionised, thus liberating electrons into the conduction band of the crystal. The occupancy (proportion of ionised donors) can be represented by statistics [8, 9], but this is not of particular concern here as most experiments are performed at liquid helium temperatures where all of the donors can be considered occupied, or at room temperature where they can be considered ionised.

At first sight, the direct analogy of acceptors in p-type material would appear to be described by the above equations, but with the hole mass replacing the electron mass and the binding energy now referring to a hole bound to a negatively charged acceptor. The hydrogenic model for a hole bound to an acceptor would give:

$$E_{\mathrm{A}^0} = -\frac{m_{\mathrm{h}}^* e^4}{32\pi^2\hbar^2\epsilon_{\mathrm{r}}^2\epsilon_0^2} \quad \text{and} \quad \lambda = \frac{4\pi\epsilon_{\mathrm{r}}\epsilon_0\hbar^2}{m_{\mathrm{h}}^* e^2}$$

Taking the typical values for GaAs of $m_{\mathrm{hh}}^* = 0.62\, m_0$ and $\epsilon_{\mathrm{r}} = 13.18$ again, then:

$$E_{\mathrm{A}^0} = -49.0\,\mathrm{meV} \quad \text{and} \quad \lambda = 11\,\text{Å}$$

Unlike the donor case, this does not agree well with experimentally measured values [5]. In fact, in practice the acceptor state is much more complex for a number of reasons. First, often the valence band at the Γ minimum consists of two degenerate states (the light and heavy holes), and thus it is not clear which effective mass should be employed, or indeed whether the hole is a mixture of both light- and heavy-hole states. Although, however, this degeneracy is usually broken within a quantum well, due to the differing effect of the confining potential on the effective masses. Secondly, the much larger heavy-hole mass has the effect of producing a much smaller Bohr radius λ, i.e. the hole orbits much more closely to the central Coulomb potential, as can be seen from the calculated radius above. Consequently, the approximation that the bulk relative permittivity ϵ describes the electromagnetic response of the lattice is questionable, although a technique for accounting for this problem will be introduced later, in Section 5.6. The problem of point defects will be revisited much later, in Chapter 15, and dealt with by a microscopic model, which will take these effects into account. For now concentration will be focused on donors.

5.2 Binding energy in a heterostructure

When a donor is placed within a quantum well structure, the situation is considerably more complex than in the bulk, due to two additional degrees of freedom. First, the binding energy depends upon the confining potential due to the quantum well structure. In its simplest form

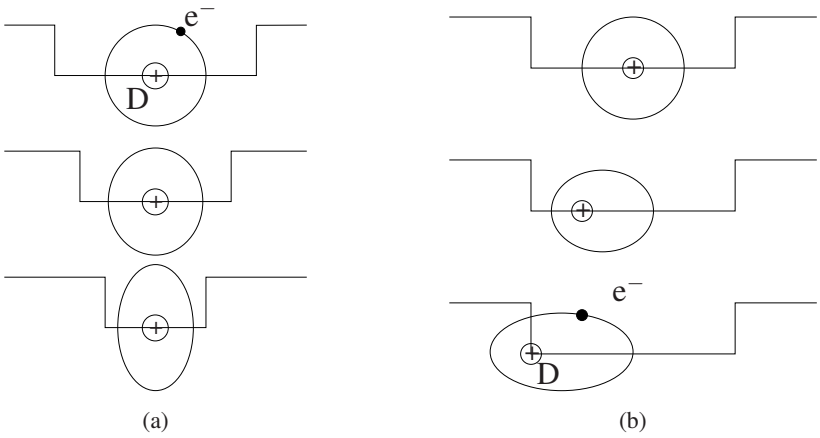


Figure 5.2: Schematic representation of the effect on the donor wave function of (a) quantum well width and (b) donor position

this would be the well width, schematically represented in Fig. 5.2(a). Beyond this, however, lies the possibility of a more complex heterostructure. The donor binding energy will be different, e.g. in a double quantum well, a superlattice and a diffused quantum well; in fact, it is just as sensitive to structure as a lone electron is.

Secondly, the donor binding energy and wave function are also a function of the donor position *within* the heterostructure. The binding energy is, of course, different for a donor at the centre of a quantum well than it is for a donor at the edge of a well, as illustrated in Fig. 5.2(b).

Any theoretical study of the properties of neutral donors in a heterostructure necessitates solving the standard Schrödinger equation for the particular structure and with the inclusion of the additional Coulombic term representing the donor potential. Within the envelope function and effective mass approximations, the Hamiltonian for an electron confined in a heterostructure and in the presence of a shallow donor is merely the standard Hamiltonian of earlier, plus an additional term due to the Coulombic interaction as follows:

$$\mathcal{H} = -\frac{\hbar^2}{2m^*}\nabla^2 + V(z) - \frac{e^2}{4\pi\epsilon r'} \tag{5.4}$$

In the interests of generality, the potential $V(z)$ describing the conduction-band-edge potential of the heterostructure will remain unspecified. The displacement between the electron and the donor is given by:

$$r'^2 = x^2 + y^2 + (z - r_\mathrm{d})^2 = x^2 + y^2 + z'^2 \tag{5.5}$$

where r_d is the position of the donor along the growth (z) direction. Note that the origin of the x–y plane has been defined on the donor atom for convenience. Note also that for now the effective mass m^* has been assumed to be constant, which leads to simplifications

in the analysis. The increased localisation of the electron within the well ensures that this approximation has little effect on the final numerical values. The extension to include the effects of a varying effective mass will be discussed later.

Methods of solution of the Schrödinger equation have centred around two basic approaches. The first of these involves expanding the electron wave function as a linear combination of Gaussian functions [10, 11]. While this technique has been successful in calculating the properties of donors in simple quantum well structures, the generalisation to more complex structures, including graded gap materials and systems where piezoelectric fields are present, is non-trivial.

The second category of approach is based on the *variational principle*. In this method a trial wave function is chosen whose functional form may contain one or more unknown parameters. These parameters are varied systematically and the expectation value of the energy calculated for each set. The variational principle [3, 4] states that *the lowest energy obtained is the closest approximation to the true state of the system*.

The success of variational approaches centres around the general choice of the trial wave function. A common choice is a product of two terms [12], i.e.

$$\Psi = \psi(z) \exp\left(-\frac{r'}{\lambda}\right) \tag{5.6}$$

where r' is the electron–donor separation and λ is a variational parameter; thus the second factor is a simple hydrogenic wave function. The function $\psi(z)$ is the uncorrelated eigenfunction, i.e. the straightforward wave function calculated in Chapters 2 and 3, of the electron in the quantum well *without* the donor [13, 14].

The latter restriction can be removed and a much more general choice of the donor wave function Ψ can be made. The motivation being generality, the problem is recast in a form which is suitable for numerical solution as described in Chapter 3, thus making it applicable to any quantum well structure, be it a double quantum well, a diffused quantum well, a graded gap quantum well, etc. With this aim, the trial wave function to be employed here is influenced by the above, but for now will be written as follows:

$$\Psi = \chi(z)\xi(x, y, z - r_{\mathrm{d}}) \tag{5.7}$$

The one-dimensional envelope function $\chi(z)$ is yet to be determined, and the function $\xi(x, y, z - r_{\mathrm{d}})$ is expected to be a hydrogenic type factor which is a function of the spatial coordinates and one or more variational parameters. It can be seen that forming the Schrödinger equation with the Hamiltonian of equation (5.4) and the trial wave function Ψ of equation (5.7) will lead to a term $\nabla^2\Psi$; it is worthwhile to derive this now. Consider:

$$\frac{\partial^2\Psi}{\partial x^2} = \frac{\partial}{\partial x}\frac{\partial}{\partial x}\left[\chi(z)\xi(x, y, z - r_{\mathrm{d}})\right] \tag{5.8}$$

$$\therefore \frac{\partial^2\Psi}{\partial x^2} = \frac{\partial}{\partial x}\left[\chi(z)\frac{\partial\xi}{\partial x}\right] = \chi\frac{\partial^2\xi}{\partial x^2} \tag{5.9}$$

and similarly for y. Consider now:

$$\frac{\partial^2\Psi}{\partial z^2} = \frac{\partial}{\partial z}\left(\frac{\partial\chi}{\partial z}\xi + \chi\frac{\partial\xi}{\partial z}\right) \tag{5.10}$$

$$\therefore \frac{\partial^2 \Psi}{\partial z^2} = \frac{\partial^2 \chi}{\partial z^2}\xi + 2\frac{\partial \chi}{\partial z}\frac{\partial \xi}{\partial z} + \chi\frac{\partial^2 \xi}{\partial z^2} \tag{5.11}$$

Hence

$$\nabla^2 \Psi = \left(\nabla_z^2 \chi\right)\xi + 2\nabla_z \chi \nabla_z \xi + \chi \nabla^2 \xi \tag{5.12}$$

where the subscript z implies derivatives with respect to z only. Then forming the Schrödinger equation with the Hamiltonian of equation (5.4) gives the following:

$$-\frac{\hbar^2}{2m^*}\left\{[\nabla_z^2 \chi(z)]\xi + 2\nabla_z \chi(z)\nabla_z \xi + \chi(z)\nabla^2 \xi\right\} - \frac{e^2}{4\pi\epsilon r'}\chi(z)\xi + V(z)\chi(z)\xi = E\chi(z)\xi \tag{5.13}$$

where E is the total energy of the system. The donor binding energy E_{D^0} is equal to the difference between this and the standard confinement energy of the electron in the heterostructure without the donor present, i.e. using the notation of Chapter 2, the ground state energy for the electron in the well without the donor would be E_1, and hence:

$$E_{\mathrm{D}^0} = E - E_1 \tag{5.14}$$

In order to proceed further, it is necessary to be more specific about the form of the hydrogenic factor ξ. In a free hydrogen atom the ground state wave function would be spherically symmetric and given by:

$$\exp\left(-\frac{r}{\lambda}\right), \qquad \text{where } r^2 = x^2 + y^2 + z^2 \tag{5.15}$$

where λ is known as the Bohr radius. Indeed, this is the form also used for donors in bulk materials. However, in heterostructures there is a loss of translational symmetry along the growth (z-)axis and hence it cannot be assumed that the spatial coordinate is spherically symmetric. For now, the hydrogenic term can be taken as:

$$\xi(x, y, z - r_{\mathrm{d}}) = \exp\left(-\frac{r''}{\lambda}\right) \tag{5.16}$$

where the spatial coordinate r'' is, of course, a function of x, y and $z - r_{\mathrm{d}}$. Again, λ is referred to as the Bohr radius, but now it will be employed as a variational parameter in order to minimise the total energy of the system. Substituting for ξ into equation (5.13) gives:

$$-\frac{\hbar^2}{2m^*}\left\{[\nabla_z^2 \chi(z)]\exp\left(-\frac{r''}{\lambda}\right) + 2\nabla_z \chi(z)\nabla_z \exp\left(-\frac{r''}{\lambda}\right) + \chi(z)\nabla^2 \exp\left(-\frac{r''}{\lambda}\right)\right\}$$
$$-\frac{e^2}{4\pi\epsilon r'}\chi(z)\exp\left(-\frac{r''}{\lambda}\right) + V(z)\chi(z)\exp\left(-\frac{r''}{\lambda}\right) = E\chi(z)\exp\left(-\frac{r''}{\lambda}\right) \tag{5.17}$$

Multiplying by $\exp(-r''/\lambda)$ and integrating over the x–y plane leads to an equation of the form:

$$-\frac{\hbar^2}{2m^*}\left[\nabla_z^2 \chi(z) I_1 + 2\nabla_z \chi(z) I_2 + \chi(z) I_3\right]$$
$$-\frac{e^2}{4\pi\epsilon}\chi(z) I_4 + V(z)\chi(z) I_1 = E\chi(z) I_1 \tag{5.18}$$

where the integrals I_j ($j = 1, 2, 3, 4$) are defined as follows:

$$I_1 = \int_0^\infty \int_0^\infty \exp\left(-\frac{2r''}{\lambda}\right) \, \mathrm{d}x \, \mathrm{d}y \tag{5.19}$$

$$I_2 = \int_0^\infty \int_0^\infty \exp\left(-\frac{r''}{\lambda}\right) \nabla_z \exp\left(-\frac{r''}{\lambda}\right) \, \mathrm{d}x \, \mathrm{d}y \tag{5.20}$$

$$I_3 = \int_0^\infty \int_0^\infty \exp\left(-\frac{r''}{\lambda}\right) \nabla^2 \exp\left(-\frac{r''}{\lambda}\right) \, \mathrm{d}x \, \mathrm{d}y \tag{5.21}$$

$$I_4 = \int_0^\infty \int_0^\infty \frac{1}{r'} \exp\left(-\frac{2r''}{\lambda}\right) \, \mathrm{d}x \, \mathrm{d}y \tag{5.22}$$

Returning to the Schrödinger equation (5.18) then:

$$\nabla_z^2 \chi(z) I_1 + 2\nabla_z \chi(z) I_2 + \chi(z) \left\{ I_3 + \frac{2m^*}{\hbar^2} \frac{e^2}{4\pi\epsilon} I_4 - \frac{2m^*}{\hbar^2} [V(z) - E] I_1 \right\} = 0 \tag{5.23}$$

Remembering that the integrals I_1, I_2, I_3, and I_4 are just real numbers, equation (5.23) is just a linear second-order differential equation, very much like the one met in Chapter 3. The energy E can be solved *for any choice of* λ by expanding the derivatives in finite differences and forming an iterative shooting algorithm as before.

First, though, for simplicity sake, it is better to simplify equation (5.23) and rewrite it as:

$$\alpha \nabla_z^2 \chi(z) + \beta \nabla_z \chi(z) + \gamma \chi(z) = 0 \tag{5.24}$$

Then, by using the finite-difference expansions:

$$\nabla_z \chi(z) = \frac{\chi(z + \delta z) - \chi(z - \delta z)}{2\delta z} \tag{5.25}$$

and

$$\nabla_z^2 \chi(z) = \frac{\chi(z + \delta z) - 2\chi(z) + \chi(z - \delta z)}{(\delta z)^2} \tag{5.26}$$

equation (5.24) becomes:

$$\frac{\chi(z + \delta z) - 2\chi(z) + \chi(z - \delta z)}{(\delta z)^2} + \frac{\beta}{\alpha} \left[\frac{\chi(z + \delta z) - \chi(z - \delta z)}{2\delta z} \right] + \frac{\gamma}{\alpha} \chi(z) = 0 \tag{5.27}$$

Finally, gathering the terms in $\chi(z + \delta z)$ on the left-hand side gives:

$$\left(1 + \frac{\beta}{2\alpha} \delta z\right) \chi(z + \delta z) = \left(-1 + \frac{\beta}{2\alpha} \delta z\right) \chi(z - \delta z) + \left[2 - (\delta z)^2 \frac{\gamma}{\alpha}\right] \chi(z) \tag{5.28}$$

which is an iterative shooting equation and can be solved subject to the standard boundary conditions, $\chi(z) \to 0$ as $z \to \pm\infty$. In practice, this involves choosing exponential-growth starting conditions deep inside the quantum barrier and then progressively calculating $\chi(z)$ at points along the z-axis across the structure. The energy E is varied until the function $\chi(z)$

satisfies the standard boundary condition, i.e. it tends to zero at the other end of the structure. For more details, see Chapter 3.

The variational aspect of the calculation arises from the value of the unknown constant λ. This is varied systematically with the aim of minimising the total energy. The system coordinates are therefore represented by the minimum energy and the corresponding value of the Bohr radius λ.

The derivation thus far is independent of the choice of the form of the hydrogenic exponential term, $\exp(-r''/\lambda)$, and this summarises the versatility of this approach. In the following sections, the various choices that can be made for this 'orbital' term will be investigated. In addition, the power of this method is represented by the generality of the conduction-band-edge potential $V(z)$. No assumptions have been made relating to the form of this function and hence the method is applicable to all forms of $V(z)$, i.e. the formalism can be applied to any heterostructure.

5.3 Two-dimensional trial wave function

The simplest choice that can be made for the hydrogenic factor of the trial wave function for an electron confined in a quantum well structure in the presence of a positively charged donor is one that is dependent only upon an in-plane coordinate, for example, as follows:

$$r'' = \sqrt{x^2 + y^2} = r_\perp \tag{5.29}$$

Therefore:

$$\xi = \exp\left(-\frac{\sqrt{x^2+y^2}}{\lambda}\right) = \exp\left(-\frac{r_\perp}{\lambda}\right) \tag{5.30}$$

This choice replaces the spherical symmetry perhaps expected, *a priori*, with circular symmetry and is often referred to as the two-dimensional (2D) form. This is an unfortunate choice of words as, in fact, ξ has cylindrical symmetry, and hence is infinitely extended along the z-axis. This is another example of the word 'dimension' being interchanged with 'coordinate'.

Given this specific form for the spatial coordinate r'', it is possible to calculate the integrals I_j explicitly, and hence all of the terms in the Schrödinger equation (5.23), and the corresponding iterative shooting equation (5.28) will be known.

Consider first:

$$I_1 = \int_0^\infty \int_0^\infty \exp\left(-\frac{2r_\perp}{\lambda}\right) \, \mathrm{d}x \, \mathrm{d}y \tag{5.31}$$

In this case, and indeed all that follow, it makes sense to exploit the circular symmetry by transforming the integral from the Cartesian coordinates, x and y, to plane polar coordinates $r_\perp$ and θ, i.e.

$$I_1 = \int_0^{2\pi} \int_0^\infty \exp\left(-\frac{2r_\perp}{\lambda}\right) \; r_\perp \mathrm{d}r_\perp \, \mathrm{d}\theta \tag{5.32}$$

As the integrand, and all that follow, have no angular dependencies, the θ integral is trivial, i.e.

$$I_1 = 2\pi \int_0^\infty \exp\left(-\frac{2r_\perp}{\lambda}\right) \; r_\perp \mathrm{d}r_\perp \tag{5.33}$$

This can now be solved with integration by parts, remembering the standard formula:

$$\int_a^b u\frac{\mathrm{d}v}{\mathrm{d}x}\,\mathrm{d}x = [uv]_a^b - \int_a^b v\frac{\mathrm{d}u}{\mathrm{d}x}\,\mathrm{d}x \tag{5.34}$$

and that the independent variable is $r_\perp$, i.e. $x = r_\perp$. By choosing:

$$u = r_\perp \quad \text{and} \quad \frac{\mathrm{d}v}{\mathrm{d}r_\perp} = \exp\left(-\frac{2r_\perp}{\lambda}\right) \tag{5.35}$$

then:

$$\frac{\mathrm{d}u}{\mathrm{d}r_\perp} = 1 \quad \text{and} \quad v = -\frac{\lambda}{2}\exp\left(-\frac{2r_\perp}{\lambda}\right) \tag{5.36}$$

Substituting into equation (5.34) gives:

$$I_1 = 2\pi\left\{\left[-\frac{\lambda r_\perp}{2}\exp\left(-\frac{2r_\perp}{\lambda}\right)\right]_0^\infty - \int_0^\infty -\frac{\lambda}{2}\exp\left(-\frac{2r_\perp}{\lambda}\right)\,\mathrm{d}r_\perp\right\} \tag{5.37}$$

$$\therefore I_1 = 2\pi\left[\left(-\frac{\lambda r_\perp}{2} - \frac{\lambda^2}{4}\right)\exp\left(-\frac{2r_\perp}{\lambda}\right)\right]_0^\infty \tag{5.38}$$

$$\therefore I_1 = 2\pi\frac{\lambda^2}{4} \tag{5.39}$$

Next, consider I_2 as defined in equation (5.20):

$$I_2 = \int_0^\infty\int_0^\infty \exp\left(-\frac{r_\perp}{\lambda}\right)\nabla_z \exp\left(-\frac{r_\perp}{\lambda}\right)\,\mathrm{d}x\,\mathrm{d}y \tag{5.40}$$

Immediately it can be seen that with this particular choice of $r'' = r_\perp$ the derivative with respect to z is zero, i.e.

$$\nabla_z \exp\left(-\frac{r_\perp}{\lambda}\right) = 0 \tag{5.41}$$

$$\therefore I_2 = 0 \tag{5.42}$$

Now consider the third integral, as specified in equation (5.21):

$$I_3 = \int_0^\infty\int_0^\infty \exp\left(-\frac{r_\perp}{\lambda}\right)\nabla^2 \exp\left(-\frac{r_\perp}{\lambda}\right)\,\mathrm{d}x\,\mathrm{d}y \tag{5.43}$$

First it is necessary to calculate the terms arising from the ∇^2, so consider:

$$\begin{aligned}\frac{\partial}{\partial x}\exp\left(-\frac{r_\perp}{\lambda}\right) &= \frac{\partial}{\partial r_\perp}\exp\left(-\frac{r_\perp}{\lambda}\right)\frac{\partial r_\perp}{\partial x} \\ &= -\frac{1}{\lambda}\exp\left(-\frac{r_\perp}{\lambda}\right)\frac{\partial}{\partial x}\sqrt{x^2+y^2} \\ &= -\frac{x}{\lambda r_\perp}\exp\left(-\frac{r_\perp}{\lambda}\right)\end{aligned} \tag{5.44}$$

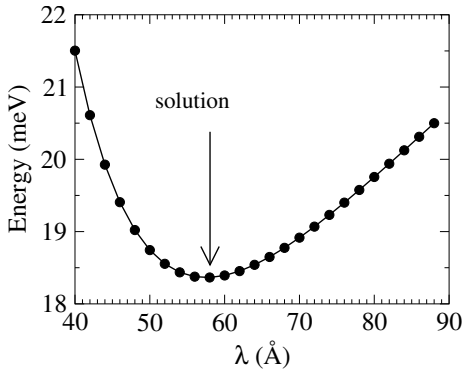


Figure 5.3: Total energy E of the electron as a function of the variational parameter λ, as an illustration of the variational principle

and then:

$$\frac{\partial^2}{\partial x^2}\exp\left(-\frac{r_\perp}{\lambda}\right) = \frac{\partial}{\partial x}\left[-\frac{x}{\lambda r_\perp}\exp\left(-\frac{r_\perp}{\lambda}\right)\right] \tag{5.45}$$

$$\therefore \frac{\partial^2}{\partial x^2}\exp\left(-\frac{r_\perp}{\lambda}\right) = -\frac{1}{\lambda r_\perp}\exp\left(-\frac{r_\perp}{\lambda}\right) - \frac{x}{\lambda}\exp\left(-\frac{r_\perp}{\lambda}\right)\frac{\partial}{\partial x}\left(\frac{1}{r_\perp}\right) - \frac{x}{\lambda r_\perp}\frac{\partial}{\partial x}\exp\left(-\frac{r_\perp}{\lambda}\right) \tag{5.46}$$

which gives:

$$\frac{\partial^2}{\partial x^2}\exp\left(-\frac{r_\perp}{\lambda}\right) = \left(-\frac{1}{\lambda r_\perp} + \frac{x^2}{\lambda r_\perp^3} + \frac{x^2}{\lambda^2 r_\perp^2}\right)\exp\left(-\frac{r_\perp}{\lambda}\right) \tag{5.47}$$

and similarly for the y-direction. Again there are no terms arising from differentiation with respect to z. Adding the two contributing terms together gives:

$$\nabla^2\exp\left(-\frac{r_\perp}{\lambda}\right) = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\exp\left(-\frac{r_\perp}{\lambda}\right) = \left(-\frac{1}{\lambda r_\perp} + \frac{1}{\lambda^2}\right)\exp\left(-\frac{r_\perp}{\lambda}\right) \tag{5.48}$$

Hence the integral I_3, as in equation (5.43), with the usual transformation to plane polar coordinates becomes:

$$I_3 = 2\pi\int_0^\infty \exp\left(-\frac{r_\perp}{\lambda}\right)\left(-\frac{1}{\lambda r_\perp} + \frac{1}{\lambda^2}\right)\exp\left(-\frac{r_\perp}{\lambda}\right)\ r_\perp \mathrm{d}r_\perp \tag{5.49}$$

Using equations (5.33) and (5.39) leads to:

$$I_3 = 2\pi\left(\int_0^\infty -\frac{1}{\lambda r_\perp}\exp\left(-\frac{2r_\perp}{\lambda}\right)\ r_\perp \mathrm{d}r_\perp + \frac{1}{\lambda^2}I_1\right) \tag{5.50}$$

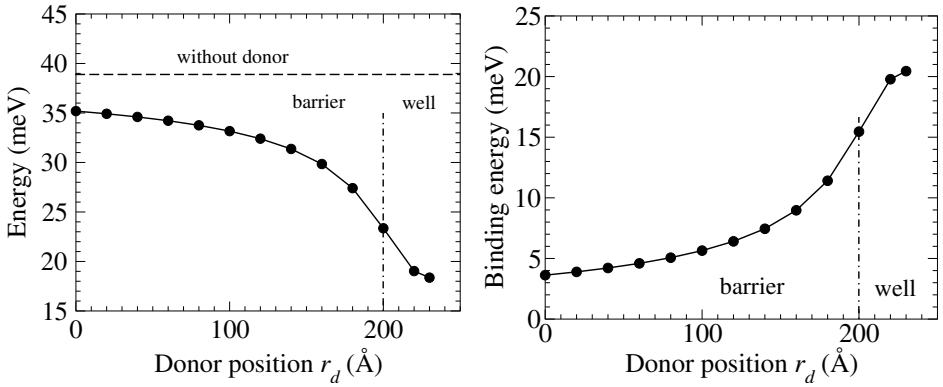


Figure 5.4: (a) Total energy E and (b) neutral donor binding energy E_{D^0} of an electron in a $60\,$Å CdTe quantum well surrounded by $Cd_{0.9}Mn_{0.1}Te$ barriers in the presence of a donor at position r_{d}

$$\therefore I_3 = 2\pi \left\{ \left[-\frac{1}{\lambda} \exp\left(-\frac{2r_\perp}{\lambda} \right) \times -\frac{\lambda}{2} \right]_0^\infty + \frac{1}{\lambda^2} I_1 \right\} \tag{5.51}$$

which gives:

$$I_3 = 2\pi \left(-\frac{1}{4} \right) \tag{5.52}$$

Finally, for the two-dimensional case, consider the only remaining integral:

$$I_4 = 2\pi \int_0^\infty \frac{1}{r'} \exp\left(-\frac{2r_\perp}{\lambda} \right) \; r_\perp \mathrm{d}r_\perp \tag{5.53}$$

Consider, for example, the substitution $r'^2 = r_\perp^2 + (z - r_{\mathrm{d}})^2 = r_\perp^2 + z'^2$. In this case the limits of integration change as follows. When $r_\perp = 0$, $r' = |z'|$ and when $r_\perp = \infty$, $r' = \infty$, hence:

$$I_4 = 2\pi \int_{|z'|}^\infty \frac{1}{r'} \exp\left(-\frac{2\sqrt{r'^2 - z'^2}}{\lambda} \right) \; r' \mathrm{d}r' \tag{5.54}$$

Note that the lower limit for r' is $|z'|$ and not just z'—this is an important point! This change of limit will be used repeatedly in this analysis. The modulus bars are present as r' is an absolute distance; it is not a vector quantity, and is therefore always positive.

Now this integration cannot be performed analytically, and it is always prudent to avoid evaluating definite integrals numerically where the upper limit is infinity. After all, exactly what value of infinity would one use? The form for I_4 can be manipulated further, however, into a more manageable integral between 0 and 1, which can then be evaluated accurately and with confidence. The mathematical trick[1] introduced here will be used again and again

[1] Credit goes to Winston Hagston for this.

in this present chapter and also later in Chapter 6. Put:

$$r' = |z'| \cosh\theta$$

Then:

$$\mathrm{d}r' = |z'| \sinh\theta \, \mathrm{d}\theta$$

and the limits become zero and infinity, which for now seems to be contrary to what was intended. Hence:

$$I_4 = 2\pi \int_0^\infty \exp\left(-\frac{2\sqrt{z'^2 \cosh^2\theta - z'^2}}{\lambda}\right) \; |z'| \sinh\theta \, \mathrm{d}\theta \tag{5.55}$$

$$\therefore I_4 = 2\pi \int_0^\infty \exp\left(-\frac{2|z'| \sinh\theta}{\lambda}\right) \; |z'| \sinh\theta \, \mathrm{d}\theta \tag{5.56}$$

Using the substitution $w = \mathrm{e}^{-\theta}$, when $\theta = 0$, $w = 1$, and when $\theta = \infty$, $w = 0$. Also, remembering that $\sinh\theta = \frac{1}{2}\left(\mathrm{e}^{\theta} - \mathrm{e}^{-\theta}\right)$, then:

$$\sinh\theta = \frac{1}{2}\left(\frac{1}{w} - w\right) \tag{5.57}$$

Equation (5.56) therefore becomes:

$$I_4 = 2\pi \int_1^0 \exp\left[-\frac{2|z'|\frac{1}{2}\left(\frac{1}{w} - w\right)}{\lambda}\right] \; |z'|\frac{1}{2}\left(\frac{1}{w} - w\right) \; \left(-\frac{\mathrm{d}w}{w}\right) \tag{5.58}$$

which gives, finally:

$$I_4 = 2\pi \int_0^1 \exp\left[-\frac{|z'|\left(\frac{1}{w} - w\right)}{\lambda}\right] \; |z'|\frac{1 - w^2}{2w^2} \, \mathrm{d}w \tag{5.59}$$

Figure 5.3 demonstrates the principle behind the variational calculation, using the above analysis, by employing a 60 Å CdTe well surrounded by 200 Å $Cd_{0.9}Mn_{0.1}Te$ barriers for illustration. The graph plots the total energy E as a function of the variational parameter λ, for an electron in the presence of a donor at the centre of the well. The system assumes the lowest energy state possible, and hence the electron moves to a Bohr orbit of radius $\lambda = 58$ Å, with energy $E = 18.366\,\mathrm{meV}$.

Figure 5.4(a) shows the results of calculations for a range of donor positions across the same 60 Å CdTe quantum well. The $Cd_{1-x}Mn_xTe$ system has been chosen to illustrate the method merely because the binding energy is larger, as calculated earlier, so it seems a bit more interesting than the more common $Ga_{1-x}Al_xAs$ system. In fact, $Cd_{1-x}Mn_xTe$ is quite similar to $Ga_{1-x}Al_xAs$; i.e. it is a direct material, with a similar band gap, and CdTe forms Type-I quantum wells with a valence-band offset similar to that of GaAs/$Ga_{1-x}Al_xAs$. None of the results obtained are particularly features of the choice of material, and re-emphasising the aim of this book, it is the theory and computational methods and the generic deductions that are important, and not the absolute values of the calculations. As mentioned already, all of

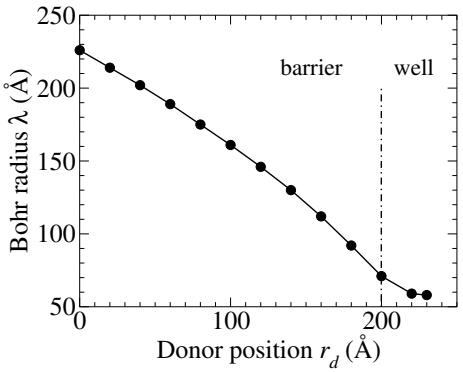


Figure 5.5: Bohr radius λ as a function of the donor position $r_{\rm d}$ across the 60 Å CdTe quantum well of above

the necessary tools are provided for the reader to quite quickly repeat the calculations for the material system of interest to them. Therefore, returning to Fig. 5.4(a), as probably expected, the electron energy E is lowest (and therefore the donor binding energy $E_{\rm D^0}$ highest) for a donor at the centre of the well. It would be expected *a priori* that the electron energy would return to the value without the donor present, as the donor is moved further and further away from the well, i.e.

$$\lim_{r_{\rm d}\to\infty} E = E_1, \qquad \text{or} \qquad \lim_{r_{\rm d}\to\infty} E_{\rm D^0} = 0 \tag{5.60}$$

This second point is highlighted by the plot of the magnitude of the binding energy, again as a function of donor position, in Fig. 5.4(b). For completeness, the corresponding plot of the variational parameter λ is displayed in Fig. 5.5. Clearly, there is a correlation between λ and $E_{\rm D^0}$, with the larger $E_{\rm D^0}$ giving the smaller λ; physically this implies that the binding energy increases as the radius of the electron orbit is decreased.

In order to produce the data plotted in Fig. 5.6, the electron ground state energies E_1 were calculated without the donor present,and with the donor present E, for a donor fixed at the centre of the well, as a function of well width. This figure illustrates two points that are both important convergence tests for this analysis and for the analysis given in Chapter 3. First, as the well width increases, the electron ground state energy without the donor present, i.e. E_1, decreases monotonically and tends towards zero. Secondly, and perhaps more importantly, and relevant to this section, as the well width increases and the contribution to the electron energy due to quantum confinement decreases, the total energy for the electron tends towards the binding energy of the neutral donor in bulk ($11.7\,\mathrm{meV}$, as calculated earlier).

The latter result is perhaps represented more clearly in Fig. 5.7. The binding energy is a non-monotonic function of the well width and peaks at a relatively small width value. This is a very similar result to that of the more complex case of exciton binding energies which will appear again in Chapter 6; in the latter case, such non-monotonic behaviour has been observed experimentally. The non-monotonic behaviour in both cases is merely due to the probability of the electron being within the well. For very narrow wells, the increasing confinement energy, as seen in Fig. 5.6, pushes the electron closer to the top of the well,

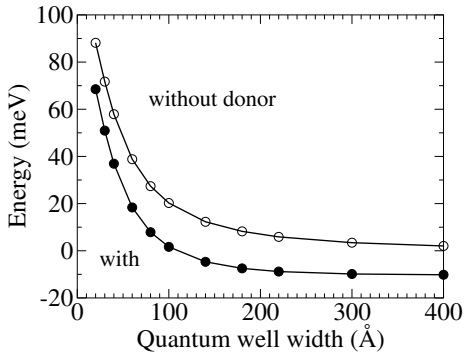


Figure 5.6: Energy of an electron in a quantum well, both with and without a donor at the centre of the well, as a function of the well width

eventually forcing it to 'spill over the top', thus leading to decreases in the binding energy. Returning to the earlier point, as the well width increases, the binding energy of the donor tends towards its bulk value, as previously hoped.

These results show that the theoretical approach is sound and, when implemented computationally, the model assumes simple limits according to sensible physical thinking. Therefore, confidence can be placed in this method. However, this form for the wave function would probably never be used in real calculations, the reason being that while it appears to be the simplest choice of trial wave function, and although it does appear to follow the expected limits, a three-dimensional trial wave function, as discussed in the next section, supersedes it. Calculations will show that a spherically symmetric wave function gives lower variational energies, which are therefore considered to be better approximations, and, in addition, the computational implementation of the mathematics is more efficient than in this, the two-dimensional case.

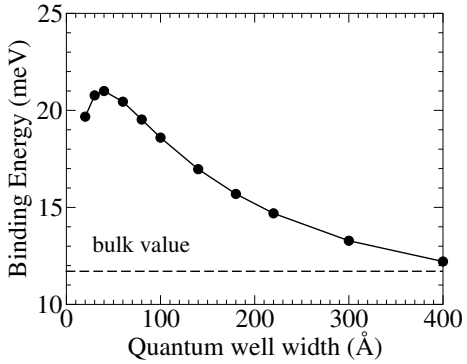


Figure 5.7: The neutral donor binding energy E_{D^0} for a donor at the centre of a well, as a function of the well width

5.4 Three-dimensional trial wave function

Driven by knowledge of bulk material and the spherical symmetry of the hydrogenic wave function, the next obvious choice of the spatial coordinate r'' is given by:

$$r'' = \sqrt{x^2 + y^2 + (z - r_{\mathrm{d}})^2} \tag{5.61}$$

which is just the electron–donor separation, defined earlier as r'. Again, the investment in the general formalism developed in Section 5.2 pays dividends in that the full requirements for the calculation merely require an evaluation of the integrals I_j ($j = 1, 2, 3, 4$), but with the new form for r''.

Consider first I_1 as originally defined in equation (5.19). Using the new form for $r'' = r'$ and switching to plane polar coordinates gives:

$$I_1 = \int_0^\infty \int_0^\infty \exp\left(-\frac{2r''}{\lambda}\right)\, \mathrm{d}x\, \mathrm{d}y = \int_0^\infty \exp\left(-\frac{2r'}{\lambda}\right)\; 2\pi\, r_\perp\, \mathrm{d}r_\perp \tag{5.62}$$

Now:

$$r'^2 = r_\perp^2 + z'^2 \tag{5.63}$$

and therefore $r'\mathrm{d}r' = r_\perp \mathrm{d}r_\perp$. Preparing for the substitution r' for $r_\perp$, when $r_\perp = 0$, $r' = |z'|$, and when $r_\perp = \infty$, $r' = \infty$, and hence:

$$I_1 = \int_{|z'|}^\infty \exp\left(-\frac{2r'}{\lambda}\right)\; 2\pi\, r'\, \mathrm{d}r' \tag{5.64}$$

which can be integrated by parts in the same manner as in Section 5.3, with the following choices:

$$u = r' \quad \text{and} \quad \frac{\mathrm{d}v}{\mathrm{d}r'} = \exp\left(-\frac{2r'}{\lambda}\right) \tag{5.65}$$

which gives:

$$\frac{\mathrm{d}u}{\mathrm{d}r'} = 1 \quad \text{and} \quad v = -\frac{\lambda}{2}\exp\left(-\frac{2r'}{\lambda}\right) \tag{5.66}$$

Thus, applying the above to equation (5.64) gives:

$$I_1 = 2\pi\left\{\left[-\frac{\lambda r'}{2}\exp\left(-\frac{2r'}{\lambda}\right)\right]_{|z'|}^\infty - \int_{|z'|}^\infty -\frac{\lambda}{2}\exp\left(-\frac{2r'}{\lambda}\right)\, \mathrm{d}r'\right\} \tag{5.67}$$

$$\therefore I_1 = 2\pi\left[-\frac{\lambda r'}{2}\exp\left(-\frac{2r'}{\lambda}\right) - \frac{\lambda^2}{4}\exp\left(-\frac{2r'}{\lambda}\right)\right]_{|z'|}^\infty \tag{5.68}$$

which after evaluation gives:

$$I_1 = 2\pi\left(\frac{\lambda|z'|}{2} + \frac{\lambda^2}{4}\right)\exp\left(-\frac{2|z'|}{\lambda}\right) \tag{5.69}$$

Turning attention to the second integral, and using the particular expression for r'' and plane polar coordinates:

$$I_2 = \int_0^\infty \int_0^\infty \exp\left(-\frac{r''}{\lambda}\right) \nabla_z \exp\left(-\frac{r''}{\lambda}\right) \, \mathrm{d}x \, \mathrm{d}y \tag{5.70}$$

$$= \int_0^\infty \exp\left(-\frac{r'}{\lambda}\right) \nabla_z \exp\left(-\frac{r'}{\lambda}\right) \; 2\pi \, r_\perp \mathrm{d}r_\perp$$

$$\therefore I_2 = 2\pi \int_0^\infty \exp\left(-\frac{r'}{\lambda}\right) \frac{\partial}{\partial r'} \exp\left(-\frac{r'}{\lambda}\right) \frac{\partial r'}{\partial z'} \; r_\perp \mathrm{d}r_\perp \tag{5.71}$$

Now:

$$r' = \sqrt{x^2 + y^2 + (z - r_\mathrm{d})^2} = \sqrt{x^2 + y^2 + z'^2} \tag{5.72}$$

and therefore:

$$\frac{\partial r'}{\partial z'} = \frac{1}{2\sqrt{x^2 + y^2 + z'^2}} \times 2z' = \frac{z'}{r'} \tag{5.73}$$

Note here the z' and *not* $|z'|$. Using this result in equation (5.71):

$$I_2 = 2\pi \int_0^\infty \exp\left(-\frac{r'}{\lambda}\right) \left(-\frac{z'}{r'\lambda}\right) \exp\left(-\frac{r'}{\lambda}\right) \; r_\perp \mathrm{d}r_\perp \tag{5.74}$$

but $r'^2 = r_\perp^2 + z'^2$, and therefore $r'\mathrm{d}r' = r_\perp \mathrm{d}r_\perp$. Using this substitution and changing the limits of integration leads to:

$$I_2 = 2\pi \int_{|z'|}^\infty \left(-\frac{z'}{r'\lambda}\right) \exp\left(-\frac{2r'}{\lambda}\right) \; r'\mathrm{d}r' \tag{5.75}$$

$$\therefore I_2 = 2\pi \left[\frac{z'}{2} \exp\left(-\frac{2r'}{\lambda}\right)\right]_{|z'|}^\infty \tag{5.76}$$

and finally:

$$I_2 = 2\pi \left[-\frac{z'}{2} \exp\left(-\frac{2|z'|}{\lambda}\right)\right] \tag{5.77}$$

Writing I_3 in plane polar coordinates with the three-dimensional form for r'':

$$I_3 = 2\pi \int_0^\infty \exp\left(-\frac{r'}{\lambda}\right) \nabla^2 \exp\left(-\frac{r'}{\lambda}\right) \; r_\perp \, \mathrm{d}r_\perp \tag{5.78}$$

In order to proceed, it is necessary to evaluate the differential. Consider:

$$\nabla_x \exp\left(-\frac{r'}{\lambda}\right) = \frac{\partial}{\partial r'} \exp\left(-\frac{r'}{\lambda}\right) \frac{\partial r'}{\partial x} \tag{5.79}$$

which, because of the isotropy of the exponential term in this case, yields, in the same manner as in equation (5.73):

$$\nabla_x \exp\left(-\frac{r'}{\lambda}\right) = -\frac{x}{r'\lambda} \exp\left(-\frac{r'}{\lambda}\right) \tag{5.80}$$

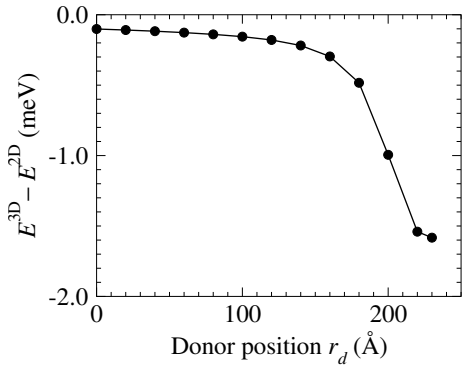


Figure 5.8: Difference in total energy for the two- and three-dimensional cases, as a function of donor position r_{d} across the 60 Å CdTe well

Differentiating again:

$$\nabla_x^2 \exp\left(-\frac{r'}{\lambda}\right) = -\frac{1}{r'\lambda}\exp\left(-\frac{r'}{\lambda}\right) - \frac{x}{\lambda}\times -\frac{1}{r'^2}\frac{\partial r'}{\partial x}\exp\left(-\frac{r'}{\lambda}\right) - \frac{x}{r'\lambda}\times -\frac{1}{\lambda}\exp\left(-\frac{r'}{\lambda}\right)\frac{\partial r'}{\partial x} \tag{5.81}$$

$$\therefore \nabla_x^2 \exp\left(-\frac{r'}{\lambda}\right) = \left(-\frac{1}{r'\lambda} + \frac{x^2}{r'^3\lambda} + \frac{x^2}{r'^2\lambda^2}\right)\exp\left(-\frac{r'}{\lambda}\right) \tag{5.82}$$

and similarly for $\nabla_y^2 \exp(-r'/\lambda)$. The same is also true for $\nabla_z^2 \exp(-r'/\lambda)$ and can be followed through by noting, however, that:

$$\nabla_z^2 \exp\left(-\frac{r'}{\lambda}\right) = \nabla_{z'}^2 \exp\left(-\frac{r'}{\lambda}\right) \times \left(\frac{\partial z'}{\partial z}\right)^2, \quad \text{where } \frac{\partial z'}{\partial z} = 1 \tag{5.83}$$

Gathering all of the terms together:

$$\nabla^2 \exp\left(-\frac{r'}{\lambda}\right) = \left(\nabla_x^2 + \nabla_y^2 + \nabla_z^2\right)\exp\left(-\frac{r'}{\lambda}\right) = \left(-\frac{2}{r'\lambda} + \frac{1}{\lambda^2}\right)\exp\left(-\frac{r'}{\lambda}\right) \tag{5.84}$$

Using this form in equation (5.78):

$$I_3 = 2\pi \int_0^\infty \exp\left(-\frac{r'}{\lambda}\right)\left(-\frac{2}{r'\lambda} + \frac{1}{\lambda^2}\right)\exp\left(-\frac{r'}{\lambda}\right)\ r_\perp\ \mathrm{d}r_\perp \tag{5.85}$$

which, using equation (5.64), gives the following:

$$I_3 = \frac{I_1}{\lambda^2} + 2\pi \int_0^\infty \left(-\frac{2}{r'\lambda}\right)\exp\left(-\frac{2r'}{\lambda}\right)\ r_\perp\ \mathrm{d}r_\perp \tag{5.86}$$

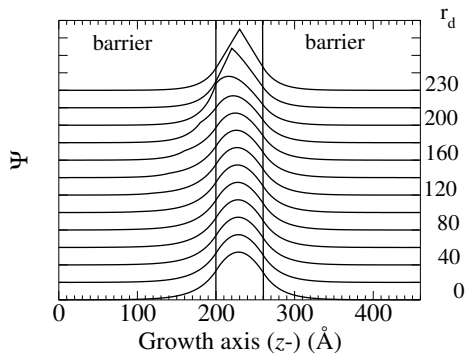


Figure 5.9: The electron wave function Ψ, as in equation (5.7), for the donor positions given on the right-hand axis

Again, substituting r' for $r_\perp$ then gives:

$$I_3 = \frac{I_1}{\lambda^2} + 2\pi\left(-\frac{2}{\lambda}\right)\int_{|z'|}^{\infty}\frac{1}{r'}\exp\left(-\frac{2r'}{\lambda}\right)\ r'\,\mathrm{d}r' \tag{5.87}$$

$$\therefore I_3 = \frac{I_1}{\lambda^2} + 2\pi\left(-\frac{2}{\lambda}\right)\left[-\frac{\lambda}{2}\exp\left(-\frac{2r'}{\lambda}\right)\right]_{|z'|}^{\infty} \tag{5.88}$$

Using the final form for I_1, as in equation (5.69):

$$I_3 = 2\pi\left(\frac{|z'|}{2\lambda} - \frac{3}{4}\right)\exp\left(-\frac{2|z'|}{\lambda}\right) \tag{5.89}$$

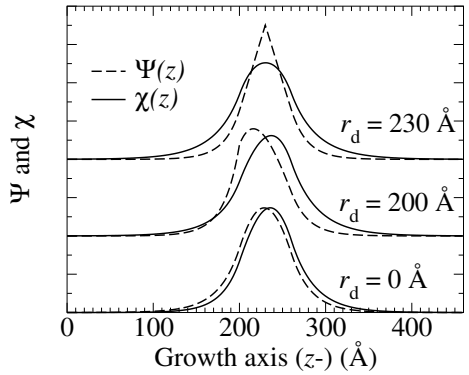


Figure 5.10: Comparison between the total electron wave function $\Psi(z)$ and the numerically determined envelope $\chi(z)$ for three different donor positions

Finally, for the three-dimensional case, I_4 becomes:

$$I_4 = 2\pi \int_0^\infty \frac{1}{r'} \exp\left(-\frac{2r'}{\lambda}\right) \; r_\perp \, \mathrm{d}r_\perp \tag{5.90}$$

which, on changing the variable to r', becomes trivial, i.e.

$$I_4 = 2\pi \int_{|z'|}^\infty \frac{1}{r'} \exp\left(-\frac{2r'}{\lambda}\right) \; r' \, \mathrm{d}r' \tag{5.91}$$

$$\therefore I_4 = 2\pi \left[\frac{\lambda}{2} \exp\left(-\frac{2|z'|}{\lambda}\right)\right] \tag{5.92}$$

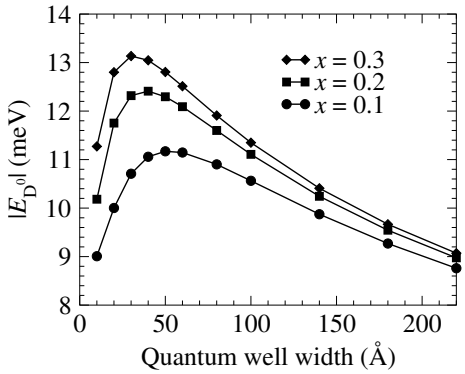


Figure 5.11: Magnitude of the donor binding energy E_{D^0} in GaAs/$\mathrm{Ga}_{1-x}\mathrm{Al}_x\mathrm{As}$ as a function of quantum well width, for a variety of barrier alloys x

Looking at these results from a computational viewpoint, it can then be seen that this 3D trial wave function has an immediate advantage over the previous, seemingly simpler, 2D case, namely that all of the integrals I_1, I_2, I_3 and I_4 have analytical expressions. Indeed, evaluation of the integrals and the minimisation of the energy is computationally much less demanding than previously.

Figure 5.8 shows the change in the total energy E of the electron between the previous 2D trial wave function and the more complex 3D case presented in this section. It is clear that the energy E is lower for *all* donor positions across this (typical) quantum well. The difference between the two trial wave functions is smallest when the separation between the donor and the electron is larger, i.e. when the donor is deep in the barrier ($r_\mathrm{d} = 0\,\text{Å}$) and the electron is, as always, centred in the well ($z = 230\,\text{Å}$). Thus it might be concluded that the 2D wave function is a reasonable approximation when the donor is in the barrier, although as the graph shows, for donors in the well, considerably lower energies can be obtained by using a spherical hydrogenic term. Recalling the variational principle, the lower energies obtained imply that the 3D approximation to the wave function is a more accurate representation than

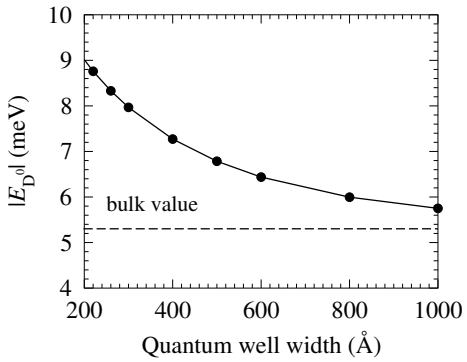


Figure 5.12: Magnitude of the donor binding energy E_{D^0} for donors at the centre of GaAs wells of large width, surrounded by $Ga_{0.9}Al_{0.1}As$ barriers

the 2D case. When coupled together with the computational advantage, as mentioned above, the argument in favour of the 3D trial wave function is clear.

Figure 5.9 displays the total wave function, $\Psi = \chi(z)\exp(-r'/\lambda)$, for the range of donor positions across the quantum well. It can be seen that the wave function Ψ resembles the one-particle wave function ψ for an electron without a donor present, for donors in the barrier, i.e. $r_{\mathrm{d}} \lesssim 160\,\text{Å}$. As the donor approaches the electron wave function, i.e. nears the barrier–well interface at $z = 200\,\text{Å}$, the electron is drawn distinctly to the left towards the donor. The influence of the hydrogenic factor $\exp(-r'/\lambda)$ can be seen for the donor positions $r_{\mathrm{d}} = 220$ and $230\,\text{Å}$ within the well.

In the 2D case, the total wave function was given by:

$$\Psi(z) = \chi(z)\exp\left(-\frac{\sqrt{x^2+y^2}}{\lambda}\right) \tag{5.93}$$

and hence the z-dependence is merely $\Psi(z) = \chi(z)$. Furthermore, it was found that the numerically determined envelope $\chi(z)$ was a very close approximation to the electron wave function $\psi(z)$ without the donor present. In this case, as is clear from Fig. 5.9, $\Psi(z) \neq \chi(z)$; this is illustrated more clearly in Fig. 5.10.

Moving on to the GaAs/$Ga_{1-x}Al_xAs$ material system and employing the bulk values of $m^* = 0.067m_0$ and $\epsilon = 13.18$, Fig. 5.11 shows the effect of well width on the neutral donor binding energy, for donors at the centre of the well, for a variety of barrier compositions. As would be expected from earlier results, E_{D^0} peaks at a narrow well width and then tails off towards the bulk value. This important limit is explored further in Fig. 5.12; it is clear from this figure that the convergence is very close, and this helps give justification to the methods developed. The variation in the Bohr radius is displayed in Fig. 5.13, and it too converges well to a value of $104\,\text{Å}$ at very large well widths. This compares admirably with the value deduced from the simple bulk hydrogenic model of $103\,\text{Å}$, calculated at the beginning of the chapter. The very small difference could arise from the finite λ increment employed of $1\,\text{Å}$.

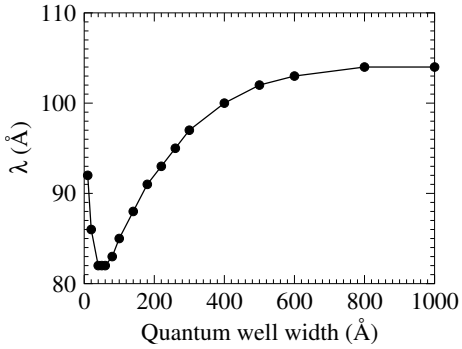


Figure 5.13: The Bohr radius λ as a function of well width for donors at the centres of GaAs wells surrounded by $Ga_{0.9}Al_{0.1}As$ barriers

5.5 Variable-symmetry trial wave function

In the two previous sections, simple high-symmetry trial wave functions have been chosen to illustrate the method. Indeed, often they provide approximate, but quick, numerical results that may generally suffice. However, a little thought highlights their deficiencies. The aim of this work, like the solution of many physical systems with the variational principle, is to choose a trial wave function that is applicable no matter what the system parameters may be—which in this case would mean no matter what the well width, barrier height and donor position are.

The spherical 3D trial wave function was chosen as this is the form found in bulk materials, i.e. systems which do not have a confining potential. Hence it might be expected that this trial wave function is most appropriate for systems which most closely resemble bulk systems, i.e. those with only small confinement, such as wide wells and/or low barrier heights. For example, for a donor in the centre of a wide well, the envelope $\chi(z)$ would look like $\cos kz$, which would be slowly varying, and hence the total wave function Ψ would resemble the bulk. Conversely, a large confinement potential might be expected to give a more 2D-like wave function.

It is clear, therefore, that while the two trial wave functions considered so far are useful, a more general form can be taken for the hydrogenic factor, a variable-symmetry term, which can move between the 2D and 3D forms should it prove energetically favourable to do so. The following equation summarises the most general choice for the hydrogenic factor of the trial wave function:

$$r'' = \sqrt{x^2 + y^2 + \zeta^2 (z - r_{\mathrm{d}})^2} \tag{5.94}$$

where ζ is a second variational parameter, which like λ will be adjusted systematically in order to minimise the total energy E of the system. Hence:

$$\xi = \exp\left(-\frac{\sqrt{x^2 + y^2 + \zeta^2 z'^2}}{\lambda} \right) \tag{5.95}$$

The effect of the parameter ζ on the symmetry of ξ is illustrated in Fig. 5.14.

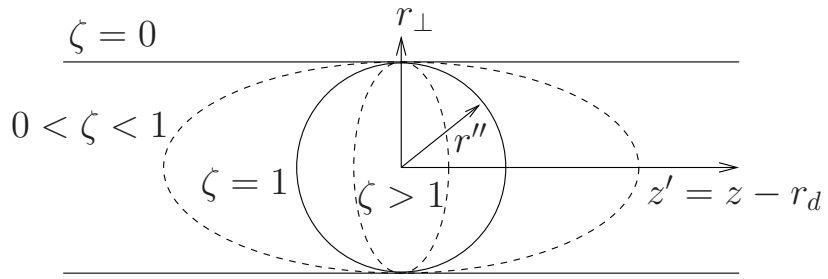


Figure 5.14: Schematic illustration of the variable-symmetry relative motion term in the donor trial wave function

Again, the investment made earlier, in structuring the problem in simple units, means that all that needs to be done in order to specify all of the quantities in the differential equation (5.23) exactly is to evaluate the integrals I_j with the new form for r''. Consider I_1 in the usual plane polar coordinates, i.e.

$$I_1 = 2\pi \int_0^\infty \exp\left(-\frac{2r''}{\lambda}\right) \; r_\perp \, \mathrm{d}r_\perp \tag{5.96}$$

Now $r''^2 = r_\perp^2 + \zeta^2 z'^2$, and therefore $r'' \mathrm{d}r'' = r_\perp \mathrm{d}r_\perp$, considering the limits, i.e. when $r_\perp = 0$, $r'' = \zeta|z'|$, and when $r_\perp = \infty$, $r'' = \infty$. Therefore:

$$I_1 = 2\pi \int_{\zeta|z'|}^\infty \exp\left(-\frac{2r''}{\lambda}\right) \; r'' \, \mathrm{d}r' \tag{5.97}$$

As previously, this can be integrated by parts. Choosing:

$$u = r'' \quad \text{and} \quad \frac{\mathrm{d}v}{\mathrm{d}r''} = \exp\left(-\frac{2r''}{\lambda}\right)$$

then:

$$\frac{\mathrm{d}u}{\mathrm{d}r''} = 1 \quad \text{and} \quad v = -\frac{\lambda}{2}\exp\left(-\frac{2r''}{\lambda}\right)$$

Therefore:

$$I_1 = 2\pi \left\{ \left[-\frac{r''\lambda}{2} \exp\left(-\frac{2r''}{\lambda}\right) \right]_{\zeta|z'|}^\infty - \int_{\zeta|z'|}^\infty \left(-\frac{\lambda}{2}\right) \exp\left(-\frac{2r''}{\lambda}\right) \; \mathrm{d}r'' \right\} \tag{5.98}$$

and so:

$$I_1 = 2\pi \left[\left(-\frac{r''\lambda}{2} - \frac{\lambda^2}{4} \right) \exp\left(-\frac{2r''}{\lambda}\right) \right]_{\zeta|z'|}^\infty \tag{5.99}$$

leading finally to:

$$I_1 = 2\pi \left(\frac{\zeta|z'|\lambda}{2} + \frac{\lambda^2}{4} \right) \exp\left(-\frac{2\zeta|z'|}{\lambda}\right) \tag{5.100}$$

Secondly:

$$I_2 = 2\pi \int_0^\infty \exp\left(-\frac{r''}{\lambda}\right) \nabla_z \exp\left(-\frac{r''}{\lambda}\right) \; r_\perp \, \mathrm{d}r_\perp \tag{5.101}$$

In order to proceed, the differential must be evaluated, i.e.

$$\nabla_z \exp\left(-\frac{r''}{\lambda}\right) = \frac{\partial}{\partial z} \exp\left(-\frac{r''}{\lambda}\right) = \frac{\partial}{\partial z''} \exp\left(-\frac{r''}{\lambda}\right) \frac{\partial z''}{\partial z}, \qquad \text{where } z'' = \zeta z' \tag{5.102}$$

With this final substitution, i.e. the use of z'', this form resembles the 3D version from earlier, and use can be made of some of the results. For example, using equation (5.80):

$$\nabla_z \exp\left(-\frac{r''}{\lambda}\right) = -\frac{z''}{r''\lambda} \exp\left(-\frac{r''}{\lambda}\right) \frac{\partial z''}{\partial z} \qquad \left(\text{note that } \frac{\partial z''}{\partial z} = \zeta\right) \tag{5.103}$$

Therefore, equation (5.101) becomes:

$$I_2 = 2\pi \int_0^\infty \exp\left(-\frac{r''}{\lambda}\right) \left(-\frac{\zeta z''}{r''\lambda}\right) \exp\left(-\frac{r''}{\lambda}\right) \; r_\perp \, \mathrm{d}r_\perp \tag{5.104}$$

Returning to z' and substituting r'' for $r_\perp$ yields:

$$I_2 = 2\pi \int_{\zeta|z'|}^\infty \left(-\frac{\zeta^2 z'}{r''\lambda}\right) \exp\left(-\frac{2r''}{\lambda}\right) \; r'' \, \mathrm{d}r'' \tag{5.105}$$

As the values r'' cancel, the integration then becomes trivial, and indeed:

$$I_2 = 2\pi \left(-\frac{\zeta^2 z'}{2}\right) \exp\left(-\frac{2\zeta|z'|}{\lambda}\right) \tag{5.106}$$

In order to evaluate I_3, it is again necessary to deduce $\nabla^2 \exp(-r''/\lambda)$ first. In a similar manner to equation (5.102):

$$\nabla_z^2 \exp\left(-\frac{r''}{\lambda}\right) = \nabla_{z''}^2 \exp\left(-\frac{r''}{\lambda}\right) \left(\frac{\partial z''}{\partial z}\right)^2 = \zeta^2 \nabla_{z''}^2 \exp\left(-\frac{r''}{\lambda}\right) \tag{5.107}$$

Considering z'' as the direct analogy of z' for the 3D case, use can be made of the earlier result in equation (5.82), hence:

$$\nabla_z^2 \exp\left(-\frac{r''}{\lambda}\right) = \zeta^2 \left(-\frac{1}{r''\lambda} + \frac{z''^2}{r''^3\lambda} + \frac{z''^2}{r''^2\lambda^2}\right) \exp\left(-\frac{r''}{\lambda}\right) \tag{5.108}$$

Using the equivalent forms for $\nabla_x^2 \exp(-r''/\lambda)$ and $\nabla_y^2 \exp(-r''/\lambda)$, also given in equation (5.82), and summing gives:

$$\nabla^2 \exp\left(-\frac{r''}{\lambda}\right) = \tag{5.109}$$
$$\left[\frac{-1-\zeta^2}{r''\lambda} + \frac{1}{\lambda^2} + \frac{(\zeta^2-1)z''^2}{r''^3\lambda} + \frac{(\zeta^2-1)z''^2}{r''^2\lambda^2}\right] \exp\left(-\frac{r''}{\lambda}\right)$$

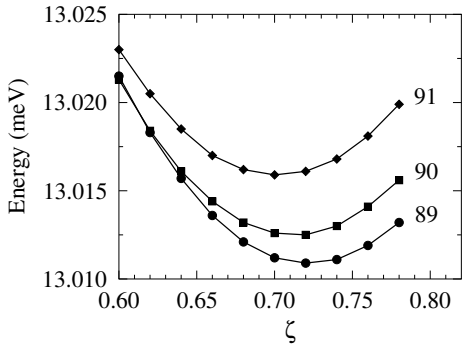


Figure 5.15: Total energy E of an electron bound to a donor as a function of the variable-symmetry parameter ζ for $\lambda = 89$, 90 and 91 Å, thus illustrating the convergence of the two-parameter variational calculation

Noting that $z'' = \zeta z'$, then:

$$\nabla^2 \exp\left(-\frac{r''}{\lambda}\right) = \left[\frac{-1-\zeta^2}{r''\lambda} + \frac{1}{\lambda^2} + \frac{(\zeta^4-\zeta^2)z'^2}{r''^3\lambda} + \frac{(\zeta^4-\zeta^2)z'^2}{r''^2\lambda^2}\right] \exp\left(-\frac{r''}{\lambda}\right) \tag{5.110}$$

As a check on the analysis so far, putting $\zeta = 0$ and $r'' = r_\perp$ does indeed yield the equivalent equation for the 2D case (equation (5.48)), while putting $\zeta = 1$ and $r'' = r'$ gives the 3D case expressed in equation (5.84).

Therefore, everything is in place for the evaluation of the third integral, which now becomes:

$$I_3 = 2\pi \int_0^\infty \left[\frac{-1-\zeta^2}{r''\lambda} + \frac{1}{\lambda^2} + \frac{(\zeta^4-\zeta^2)z'^2}{r''^3\lambda} + \frac{(\zeta^4-\zeta^2)z'^2}{r''^2\lambda^2}\right] \exp\left(-\frac{2r''}{\lambda}\right) \; r_\perp \, \mathrm{d}r_\perp \tag{5.111}$$

Unfortunately each of the four terms needs to be handled separately, so equation (5.111) is written as:

$$I_3 = 2\pi \left(I_{31} + I_{32} + I_{33} + I_{34}\right) \tag{5.112}$$

Then the first integral is trivial, while the second can be solved by parts, thus giving:

$$I_{31} = \frac{-1-\zeta^2}{2} \exp\left(-\frac{2\zeta|z'|}{\lambda}\right) \quad \text{and} \quad I_{32} = \left(\frac{\zeta|z'|}{2\lambda} + \frac{1}{4}\right) \exp\left(-\frac{2\zeta|z'|}{\lambda}\right) \tag{5.113}$$

The third in this short series cannot be solved analytically; however, it can again be manipulated into an integral between 0 and 1, and hence quickly and accurately evaluated. It is a worthwhile exercise to follow this one through; changing the variable of integration from $r_\perp$ to r'' leads to:

$$I_{33} = \int_{\zeta|z'|}^\infty \left[\frac{(\zeta^4-\zeta^2)z'^2}{r''^3\lambda}\right] \exp\left(-\frac{2r''}{\lambda}\right) \; r'' \, \mathrm{d}r'' \tag{5.114}$$

Putting $r'' = \zeta|z'|\cosh\theta$, then $\mathrm{d}r'' = \zeta|z'|\sinh\theta\ \mathrm{d}\theta$, and it follows that:

$$I_{33} = \int_0^\infty \frac{(\zeta^4 - \zeta^2)z'^2}{\lambda\zeta|z'|\cosh^2\theta} \exp\left(-\frac{2\zeta|z'|\cosh\theta}{\lambda}\right) \sinh\theta\ \mathrm{d}\theta \tag{5.115}$$

Again, making the substitution $w = \mathrm{e}^{-\theta}$, and rewriting the $\sinh$ and $\cosh$ terms as in equation (5.57), yields the final form:

$$I_{33} = \int_0^1 \frac{2(\zeta^3 - \zeta)|z'|}{\lambda} \exp\left[-\frac{\zeta|z'|(\frac{1}{w} + w)}{\lambda}\right] \frac{1 - w^2}{(1 + w^2)^2}\ \mathrm{d}w \tag{5.116}$$

Similarly:

$$I_{34} = \int_0^1 \frac{2(\zeta^4 - \zeta^2)z'^2}{\lambda^2} \exp\left(-\frac{\zeta|z'|(\frac{1}{w} + w)}{\lambda}\right) \frac{1 - w^2}{w(1 + w^2)}\ \mathrm{d}w \tag{5.117}$$

Finally, I_4 follows similarly, with the variable of integration $r_\perp$ needing to be switched in order to match the denominator r', thus allowing manipulation to give:

$$I_4 = 2\pi \int_0^1 \exp\left[-\frac{2|z'|\sqrt{\left(\frac{1-w^2}{2w}\right)^2 + \zeta^2}}{\lambda}\right] |z'| \frac{1 - w^2}{2w^2}\ \mathrm{d}w \tag{5.118}$$

It should be noted that putting $\zeta = 1$ in the above integrals, I_1, I_2, I_3, and I_4, reassuringly gives the same results as those found for the 3D case, while $\zeta = 0$ gives the same results as the 2D case. The manipulation of I_4 is not trivial, however, and is left as an exercise for the reader (see exercise (4)). Numerical convergence tests have confirmed that $\zeta = 1$ does indeed yield the 3D energies, while $\zeta = 0$ gives the 2D energies, although in both cases it is necessary to use a fine mesh, i.e. a small value of δz in equation (5.28).

Figure 5.15 gives an example of the output from this two-parameter variational calculation. For the record, this calculation is for a donor at the centre of a 100 Å GaAs well surrounded by $Ga_{0.9}Al_{0.1}As$ barriers. For the sake of clarity, only the Bohr radii corresponding to the minima (at $\lambda = 89$ Å) and above are shown. It can be seen from this figure that the change in energy due to the ζ variation is very small, and certainly beyond the scope of experimental measurement. Nonetheless, the graph shows that the wave function will assume a value of $\zeta \approx 0.72$. This will have implications for systems where the extent and nature of the wave function are important, e.g. electron scattering from neutral donors, donor-bound excitons, and the electrical excitation of impurities.

Table 5.1 compares the energies obtained with all three trial wave functions for donor positions across a 60 Å CdTe well, surrounded by 200 Å $Cd_{0.9}Mn_{0.1}Te$ barriers (hence the centre of the well is at 230 Å). Note, as mentioned above, that the trial wave functions which contain integrals that require numerical evaluation, i.e. the 2D and the variable-symmetry case which contain the integrals between 0 and 1 with respect to w, demand around eight wave function points per angstrom for complete convergence. The data displayed earlier in Fig. 5.8 as a comparison between the 2D and the 3D case were obtained quickly with the standard mesh of one point per angstrom, i.e. $\delta z = 1$ Å. This was acceptable at that point in order to illustrate merely that the 3D case gave lower variational energies than the 2D case.

Table 5.1 Total energies for an electron bound to a donor as a function of donor position in a $60\,$Å CdTe quantum well surrounded by $Cd_{0.9}Mn_{0.1}Te$ barriers, for the three different trial wave functions (data first appeared in [15]; data reproduced by permission of Academic Press)

r_d (Å)	$E(\zeta = 0)$ (meV)	$E(\zeta = 1)$ (meV)	$E(\zeta)$ (meV)
150	30.72	30.48	30.44
160	29.89	29.60	29.56
170	28.83	28.48	28.44
180	27.45	27.01	26.97
190	25.62	25.05	25.02
200	23.22	22.50	22.47
210	20.49	19.71	19.67
220	18.30	17.58	17.53
230	17.47	16.80	16.74

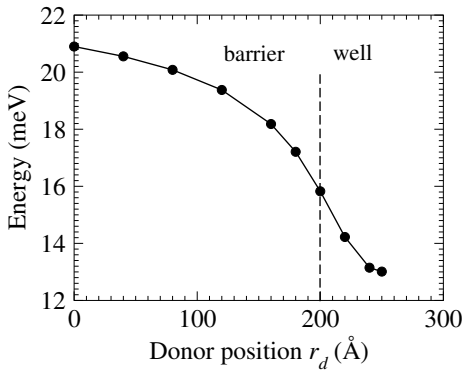


Figure 5.16: Energy E as a function of donor position across a $100\,$Å GaAs well surrounded by $Ga_{0.9}Al_{0.1}As$ barriers

However, for the purposes of demonstrating that the variable-symmetry case gives lower energies still, more accurate calculations were necessary, and indeed for Table 5.1, δz in the iterative equation (equation (5.28)) was required to be $0.1\,$Å. The disadvantage to moving to such accuracy is, of course, the increased computational effort, which is roughly inversely proportional to δz; hence an increase by a factor of 10 in the number of wave function points leads to a similar increase in computation time. Table 5.1 clearly illustrates that the variable-symmetry trial function gives the lowest energies, for all donor positions, although in terms of energy the 3D wave function is a very good approximation also, with energies differing by less than $0.1\,\mathrm{meV}$. In comparison with the more accurate evaluation of the energy for the 2D case, this differs by up to $0.7\,\mathrm{meV}$.

Figures 5.16 and 5.17 illustrate the results of detailed calculations for the GaAs/$Ga_{1-x}Al_xAs$ system. The energies in Fig. 5.16 are similar in form to those obtained earlier,

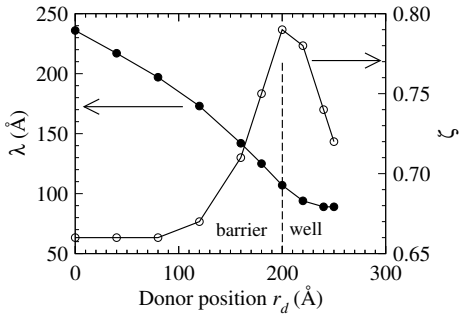


Figure 5.17: Bohr radius λ and symmetry parameter ζ as a function of donor position

with the interest here lying with the symmetry of the wave function, as given in Fig. 5.17. Most obviously, it is clear that the values of ζ imply that the donor wave function will assume symmetries that are neither 2D or 3D, thus justifying the investment in the variable-symmetry analysis. The value of ζ is non-monotonic as a function of donor position, with a local minimum for donors at the centre of the well and a peak for donors near the well–barrier interface; this is the exact same behaviour as reported by Roberts *et al*. [15] for the CdTe/$Cd_{1-x}Mn_xTe$ system and hence appears to be a generic result. Typically, it is found that $\zeta \sim 0.7$, and hence the variable-symmetry trial wave function takes the form:

$$\Psi = \chi(z) \exp\left(\frac{\sqrt{x^2 + y^2 + z'^2/2}}{\lambda}\right) \tag{5.119}$$

i.e. the coefficient of z'^2 is half that of x and y. Therefore, the hydrogenic factor is midway between the 2D and 3D case, with a very definite prolate spheroid shape (refer again to Fig. 5.14).

In conclusion, regarding the various advantages and disadvantages of the three trial wave functions, clearly the last, namely the variable-symmetry case, gives the lowest variational energies and hence is the most accurate physical representation. Without such an investment in complexity and computation time, it would not have been apparent that donors assume this mixed 2D and 3D symmetry, i.e. wave functions resembling prolate spheroids. However, the computational demands are *substantial*, and it has been shown that donor energies can be obtained within $0.1\,\mathrm{meV}$ of the lowest possible value by using a simple 3D (spherical) trial wave function. This wave function has the advantage that all of the relevant integrations can be performed analytically, which leads to a quick evaluation of the total energy E.

5.6 Inclusion of a central cell correction

As mentioned earlier, the bulk dielectric constant has to be employed with caution. In the majority of cases its use is probably fine; however, it is worthwhile noting that there is a scheme for accounting for a change in the permittivity of the material as the electron approaches the donor. This occurs as the degree of electronic shielding is reduced and is known as the *central cell correction*.

This extension takes the form of an additional factor in the Coulombic potential term, which, for the purpose of illustration, is taken to be as follows:

$$-\frac{e^2}{4\pi\epsilon r'} \longrightarrow -\frac{e^2}{4\pi\epsilon r'}\left(1+\Xi \mathrm{e}^{-\frac{r'}{\Lambda}}\right) \tag{5.120}$$

where Λ is a parameter describing the extent of the correction; the latter can be thought of as a screening length. As $r' \longrightarrow 0$, it is expected that:

$$-\frac{e^2}{4\pi\epsilon r'}\left(1+\Xi \mathrm{e}^{-\frac{r'}{\Lambda}}\right) \longrightarrow -\frac{e^2}{4\pi\epsilon_0 r'} \tag{5.121}$$

where ϵ_0 is the permittivity of free space. Taking this limit gives:

$$\frac{1}{\epsilon}(1+\Xi) = \frac{1}{\epsilon_0} \tag{5.122}$$

which when rearranged gives:

$$\Xi = \frac{\epsilon}{\epsilon_0} - 1 \tag{5.123}$$

The effect of these changes on the analysis carried out above occurs solely in the integral I_4, which for the most general case would become:

$$I_4' = 2\pi \int_0^\infty \left[1+\Xi \exp\left(-\frac{r'}{\Lambda}\right)\right] \frac{1}{r'} \exp\left(-\frac{2r''}{\lambda}\right) \; r_\perp \, \mathrm{d}r_\perp \tag{5.124}$$

If the particular 3D case were chosen, then this gives:

$$I_4' = I_4 + 2\pi\Xi \frac{\Lambda\lambda}{\lambda+2\Lambda} \exp\left(-\frac{|z'|}{\Lambda}\right) \exp\left(-\frac{|z'|}{\lambda}\right) \tag{5.125}$$

Recent work has demonstrated the effect of such a central cell correction for this 3D case [16]. While the physical reasoning behind the addition of such a correction seems sound, here, as always, caution needs to be exercised in choosing a value for Λ. Cynically, Λ could be looked upon as merely another parameter (although it is *not* a variational parameter), which can be varied in order to produce agreement between experiment and theory.

The evaluation of I_4' for the most general, variable-symmetry case will be left as an exercise for the reader.

5.7 Special considerations for acceptors

Hitherto, the focus has been on donor energies in heterostructures, although, as discussed earlier in this chapter, *in principle* all of the analysis above is also applicable to the calculation of acceptor levels. Agreement between the hydrogenic model of the acceptor in bulk and experiment needs to account for both the degeneracy of the valence band and the reduced screening of the negatively charged acceptor ion potential. In heterostructures, however, the confinement potential lifts the degeneracy of light and heavy holes, so the only additional consideration that needs to be made relates to the permittivity.

The procedure would be to deduce empirically the central cell correction parameter Λ, by comparing the calculated acceptor binding energy E_{A^0} predicted by the above analysis and adjusting Λ to give agreement with experiment, for any well-characterised quantum well system. For a particular material system and acceptor atom, the screening length Λ *should be* constant. Therefore it can be used in further predictive calculations for different well widths, barrier heights, dopant positions, etc.

Acceptor levels within quantum wells have been calculated successfully, as discussed by Bastard [17]. Masselink *et al*. [18] have achieved good agreement with the experimental measurements of Miller *et al*. [19] for the binding energy of carbon acceptors in GaAs/$Ga_{1-x}Al_xAs$ quantum wells.

5.8 Effective mass and dielectric mismatch

A great deal of attention has been paid in the literature to the role of effective mass mismatch at interfaces in semiconductor heterostructures (see Bastard [17]) and to the effect that this may have on donor energies. In addition, Fraizzoli *et al.* [20] have studied in detail the role of a dielectric constant mismatch at interfaces between dissimilar materials and its effect on shallow donor impurity levels in GaAs/$Ga_{1-x}Al_xAs$ quantum well structures. Such considerations can be readily incorporated into the present approach. At first sight it might be thought that this can be achieved by simply allowing m^* and ϵ to be functions of z in equation (5.23) and in the integrals $I_1, \ldots, I_4$. However, as discussed in detail in Chapter 3, the change to a variable effective mass necessitates substantial alterations to the quantum mechanical kinetic energy operator. Again, as before, the first term in equation (5.4) would need to be substituted with:

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial z^2} \longrightarrow -\frac{\hbar^2}{2}\frac{\partial}{\partial z}\frac{1}{m^*(z)}\frac{\partial}{\partial z} \tag{5.126}$$

In principle, the analysis could be followed through again, with re-evaluation of the integrals affected by this substitution.

The generalisation to include the material dependency of the permittivity ϵ is simpler to deal with. It can be achieved by absorbing this function into the integral I_4 which arises from the Coulombic term, i.e. with the appropriate change in the coefficient, I_4 would become:

$$I_4' = 2\pi \int_0^\infty \frac{1}{\epsilon(z)r'} \exp\left(-\frac{2r''}{\lambda}\right) \; r_\perp \, \mathrm{d}r_\perp \tag{5.127}$$

as ϵ depends on the electron position and not simply on the electron–donor separation as usual. It is clear, however, that evaluation of this integral is non-trivial. To the authors' knowledge, neither of these generalisations to the theoretical approach in this chapter has been implemented.

While some authors deem it necessary to include such dependencies in calculations [10], it must be remembered that often the interest lies in systems of weak alloys such as $Ga_{1-x}Al_xAs$ and $Cd_{1-x}Mn_xTe$ where the effective mass change between the well and barrier is relatively small. Nonetheless, the mechanisms by which the theory presented here can be extended have been mapped out, and here lies an opportunity for the interested reader to explore such systems further.

5.9 Band non-parabolicity

Small well widths and large potential barriers could require the inclusion of the non-parabolicity of the conduction band [13]. Ekenberg [21] described the inclusion of non-parabolicity on the subband structure of quantum wells. This method can account accurately for a variety of physical phenomena, but is analytically complicated. Simpler procedures have been proposed by various authors for the more complex problem of a donor in a quantum well. For example, Chaudhuri and Bajaj [13] have used the following simple replacement:

$$m^* \longrightarrow m^*(E) = a_0 + a_1 E + a_2 E^2 + a_3 E^3 + \cdots + a_n E^n \tag{5.128}$$

where a_n represents a series of constants. Given the values of these constants, the effects of non-parabolicity can be incorporated into equation (5.17) simply by making the effective mass a function of the energy E. This extension is certainly much more straightforward than the two described in the previous section.

In relation to the calculations below, it should be noted that Chaudhuri and Bajaj [13] showed that, even with relatively large potential barriers, band non-parabolicity was only significant for wells narrower than half the Bohr radius of the neutral donor, which, for the case of $Ga_{1-x}Al_xAs$, would be less than $50\,\text{Å}$, and for $Cd_{1-x}Mn_xTe$, would be less than $35\,\text{Å}$.

5.10 Excited states

Just as there are many solutions to the Schrödinger equation for an isolated hydrogen atom, there are also many more solutions representing excited energy states of a donor in a heterostructure. Recalling the hydrogen-atom solutions [2]:

$$\psi_{1\text{s}} = \exp\left(-\frac{r}{\lambda}\right), \qquad \psi_{2\text{s}} = \left(2 - \frac{r}{\lambda}\right)\exp\left(-\frac{r}{2\lambda}\right), \qquad \psi_{2\text{p}_x} = x\exp\left(-\frac{r}{2\lambda}\right) \tag{5.129}$$

and similarly for 2p_y and 2p_z. The corresponding eigenenergy involves only the ground state and the principle quantum number, i.e. $E = E_1/n^2$, so very simply $E_2 = E_1/4$, $E_3 = E_1/9$, etc. In bulk semiconductors, the neutral donor does exhibits these states (see [8], p. 314).

The situation is more complex in semiconductor heterostructures as the one-dimensional potential due to the layer structure breaks the symmetry of the spherical potential, and hence the wave function is more complex than for the hydrogen atom. Much detail has been given as to solving the corresponding ground state under these conditions. In an analogy with the hydrogen atom the first excited state of the donor might be written as follows:

$$\Psi_{2\text{s}} = \chi(z)\left(1 - \frac{\alpha r''}{\lambda_{2\text{s}}}\right)\exp\left(-\frac{r''}{\lambda_{2\text{s}}}\right) \tag{5.130}$$

where $\lambda_{2\text{s}}$ has been labelled specifically as it cannot be assumed, *a priori*, that $\lambda_{2\text{s}} = 2\lambda_{1\text{s}}$. The constant α has been introduced, and is determined by ensuring orthogonality between the ground state and the first excited state, i.e.

$$\langle \Psi_{1\text{s}} | \Psi_{2\text{s}} \rangle = \int_0^\infty \Psi_{1\text{s}}^* \Psi_{2\text{s}} \; \mathrm{d}\tau = 2\pi \int_{-\infty}^{\infty}\int_0^\infty \Psi_{1\text{s}}^* \Psi_{2\text{s}} \, \mathrm{d}r_\perp \, \mathrm{d}z = 0 \tag{5.131}$$

Studies exist in the literature of the excited states of donors in heterostructures [22], but to the authors' present knowledge the extension utilising the general form for $\Psi = \chi(z)\xi(x, y, z - r_{\mathrm{d}})$ has not yet been attempted. A simple alternative approach which yields reasonable results is presented later in this chapter.

5.11 Application to spin-flip Raman spectroscopy in diluted magnetic semiconductors

5.11.1 Diluted magnetic semiconductors

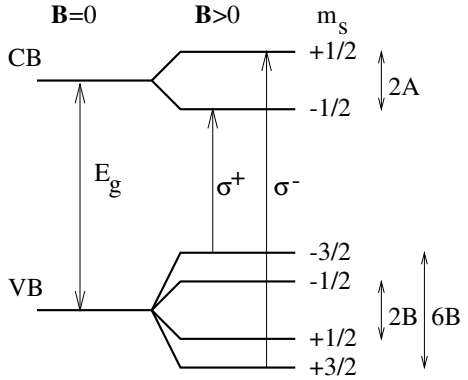


Figure 5.18: Zeeman splitting of conduction (CB) and valence (VB) bands of a diluted magnetic semiconductor within an external magnetic field **B**

Diluted magnetic semiconductors (DMSs) [23, 24] are important because of the strong exchange interaction between the hybridised sp^3–d orbitals of the magnetic ions and charge carriers. This manifests itself most clearly in the giant Zeeman splittings observed in both the conduction and valence bands when the material is placed in an external magnetic field. At low magnetic ion concentrations, the materials generally exhibit 'frustrated' paramagnetism, with the number of spin singlet states being reduced by nearest-neighbour anti-ferromagnetic spin pairing [25]. At low magnetic fields (less than $8\,\mathrm{T}$) these spins remain locked and cannot contribute to the paramagnetism of the material; however, experiments carried out under very high magnetic fields have been able to break these spin doublets. The magnetic ion itself, usually Mn^{2+}, sits substitutionally on a cation site and can generally be incorporated to high concentrations. The most common DMS is $Cd_{1-x}Mn_xTe$, while others include $Zn_{1-x}Mn_xS$, and more recently $Ga_{1-x}Mn_xAs$. Figure 5.18 shows the Zeeman effect in a DMS material, with the vertical arrows linking the heavy-hole states ($\pm 3/2$) and the electron states, thus illustrating the allowed interband transitions under circularly polarised light, i.e. $|\mathrm{hh}_{+3/2}\rangle \rightarrow |\mathrm{e}_{+1/2}\rangle$ is the σ^- transition and $|\mathrm{hh}_{-3/2}\rangle \rightarrow |\mathrm{e}_{-1/2}\rangle$ is the σ^+ transition.

The magnitudes of the conduction and valence-band splittings are given in terms of the variables A and B as follows:

$$A = -\frac{1}{6}N_0\alpha x \langle S_z \rangle , \qquad B = -\frac{1}{6}N_0\beta x \langle S_z \rangle \tag{5.132}$$

where $N_0\alpha$ and $N_0\beta$ are constants (220 and $880\,\mathrm{meV}$, respectively, in $Cd_{1-x}Mn_xTe$).[2] The expectation value of the magnetic ion spin along the z-axis $\langle S_z \rangle$ is given by:

$$\langle S_z \rangle = S_0(x) B_J(\mathbf{B}, T_{\text{eff}}) \tag{5.133}$$

where $S_0(x)$ is the effective spin of the magnetic ions and B_J is a Brillouin function describing the response of the spins in a magnetic fields **B**. The effective spin S_0 accounts for the proportion of magnetic ions which are spin-paired with a nearest neighbour and cannot respond to the alignment induced by the magnetic field. Alternatively the effective spin can be considered as the concentration of spin-singlet states, i.e.

$$xS_0(x) = \frac{5}{2}\overline{x} \tag{5.134}$$

where the spin of the Mn^{2+} ions is $5/2$. Recent theoretical studies [25] have calculated $\overline{x}$ in agreement with experiment [26] and shown that for moderate fields ($\approx 8\,\mathrm{T}$), i.e. when the splittings have saturated, but before the nearest-neighbour spin pairings are broken, the maximum value of $x \langle S_z(x) \rangle$ occurs at a manganese concentration $x \approx 0.15$ and is equal to 0.105. Hence, in $Cd_{1-x}Mn_xTe$ the maximum splitting in the conduction band is $\approx 23\,\mathrm{meV}$, and for the heavy holes in the valence band it is $\approx 92\,\mathrm{meV}$. The paramagnetic behaviour falls off with increasing temperature.

When applying a magnetic field to a semiconductor heterostructure, the direction of the field becomes important. For fields parallel to the growth (z-)axis (the Faraday configuration), the splittings are still well represented by Fig. 5.18. However, application of the magnetic field along the plane of the wells, i.e. the Voigt configuration, leads to mixing of the light- and heavy-hole valence states, thus producing a much more complex band structure. However, this lies beyond the scope of this short introduction.

Thus far, the CdTe/$Cd_{1-x}Mn_xTe$ system has been employed to illustrate donor binding-energy calculations. Now, however, specific use will be made of these magnetic properties in an application of the binding-energy calculations, namely spin-flip Raman spectroscopy.

5.11.2 Spin-flip Raman spectroscopy

The application of a magnetic field to a heterostructure leads to a Zeeman splitting of the confined energy states, where in the case of an n-type material (i.e. material with donors present), two electron states are formed with the spin-up states having a higher energy than the spin-down states. At low temperatures, the carriers populate the lowest-energy state, i.e. the spin-down, but can be excited by photons into the higher-energy state. This excitation can be detected as a Raman shift [27], as illustrated schematically in Fig. 5.19.

[2] Note that the x-variable here refers to the alloy fraction and not to the spatial direction.

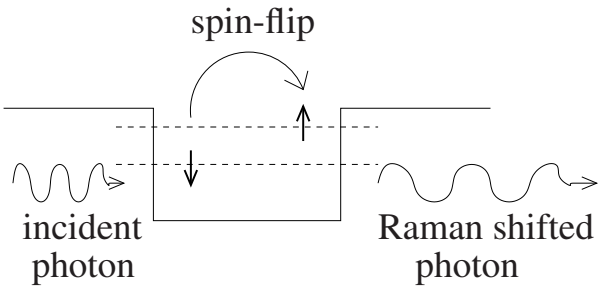


Figure 5.19: Schematic illustration of the spin-flip process

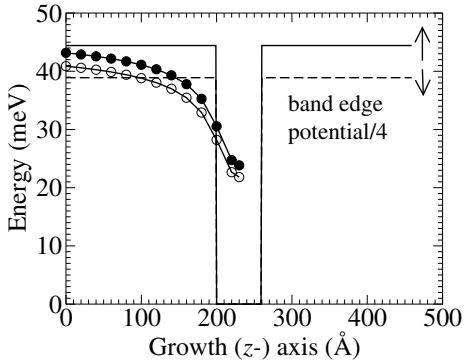


Figure 5.20: Energy of the spin-up (solid circles) and spin-down (open circles) states at a magnetic field of $8\,\mathrm{T}$, for a range of donor positions across a 60 Å CdTe well surrounded by $Cd_{0.85}Mn_{0.15}Te$ barriers

This is a very powerful technique because the spectroscopy depends only upon one carrier type and theoretical modelling of the experimental data requires the parameter set of only one band. In this case, the parameters consist of the relative permittivity of the material, and the electron effective mass from bulk, together with the conduction-band offset between the non-magnetic CdTe well and the magnetic $Cd_{1-x}Mn_xTe$ barriers [28].

The Raman shift is simply the energy difference between the two electron-donor spin states, $E_{\mathrm{sf}} = E^{\uparrow} - E^{\downarrow}$. Figure 5.20 displays the energy of these two spin states for a lightly n-doped CdTe/$Cd_{1-x}Mn_xTe$ system, calculated using the 3D trial wave function. The magnetic ion concentration in the barriers has been chosen as 15% in order to maximise the energy difference between the states. Figure 5.21 displays the spin-flip energy for this system, again as a function of donor position.

If the donors are assumed to be uniformly distributed across the entire heterostructure, it is possible to represent the data of Fig. 5.21 in the form of an intensity I versus spin-flip energy E Raman spectrum, by assigning a Gaussian distribution to each point in energy space with

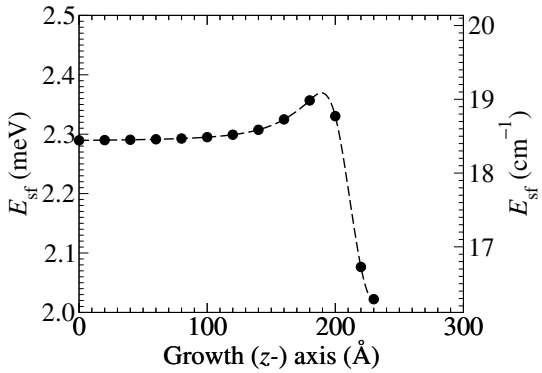


Figure 5.21: Spin-flip energy as a function of donor position

a certain linewidth, i.e.

$$I = \sum_{r_{\rm d}} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(E - E_{\rm sf}(r_{\rm d}))^2}{2\sigma^2}\right] \tag{5.135}$$

where $E_{\rm sf}(r_{\rm d})$ is the spin-flip energy of the donor at position $r_{\rm d}$. The finite linewidth of any spin-flip signal will arise from microscopic fluctuations of the material parameters. For example, the well width, alloy fluctuations in the barrier, and the random nature of the donor distribution itself will all give rise to a broadening of the signal from each of the donor positions $r_{\rm d}$ along the axis of the quantum well structure. The standard deviation σ is related to the linewidth l (full width at half maximum) by:

$$\sigma = \frac{l}{2\sqrt{2\ln 2}} \tag{5.136}$$

In order to save considerable computational effort, the donor calculations were performed only at the points marked by circles on Fig. 5.21. A spline of these data was produced, given by the dashed line, and then this more detailed curve was used to produce the intensity versus energy data of Fig. 5.22, which also shows the effect of different linewidths on the predicted Raman spectrum of this single quantum well.

In the lowest curve ($l = 0.5\,{\rm cm}^{-1}$) it is possible, in principle, to resolve the donors in the well from those in the barrier, as observed experimentally [28]. Comparing Figs. 5.21 and 5.22, it can be seen that the central peak at around $18.5\,{\rm cm}^{-1}$ corresponds to spin-flips of electrons bound to donors *in the barrier*. This peak in intensity is due entirely to the proportionately larger fraction of donors in the barrier compared with those in the well. In addition, spin-flips from electrons bound to donors in the well are clearly resolved as a small peak at about $16\,{\rm cm}^{-1}$.

As the linewidth l of the signals is allowed to increase, the resolution and information in the simulated spectrum decreases. At a linewidth of $1\,{\rm cm}^{-1}$, spin-flips due to donors in the well appear as just a shoulder on the larger 'barrier' peak, while by $l = 2\,{\rm cm}^{-1}$ this information is lost altogether, and a broad central peak ensues.

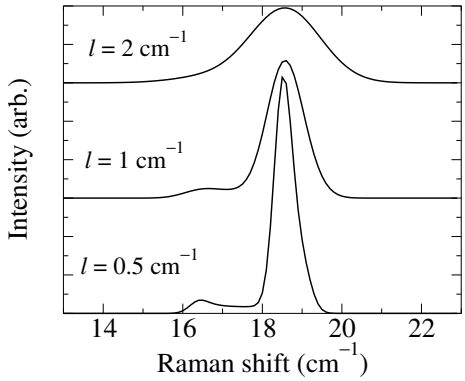


Figure 5.22: Intensity of Raman signal versus energy shift

In conclusion, the theoretical donor binding-energy calculations outlined above have been shown to be of direct relevance to simulating spin-flip Raman spectroscopy in diluted magnetic semiconductors. This allows the spatial distribution of donors to be investigated, although the importance of high-quality samples has been shown to be significant.

It has been proposed that the magnetic behaviour of the first few monolayers of a dilute magnetic semiconductor adjacent to an interface with a non-magnetic semiconductor, could be significantly different from that of the bulk [29, 30, 31]. One contribution to this effect arises from a reduction in the number of anti-ferromagnetically coupled pairs due to a decrease in the number of nearest-neighbour magnetic ions [29]. A single layer of donors, δ-doped into a quantum well structure in the region of the well–barrier interface, could provide a useful probe of the magnetism—via the observations made from spin-flip Raman spectroscopy.

5.12 Alternative approach to excited impurity states

The earlier method of calculating the energy levels of impurities within semiconductor heterostructures requires major mathematical analysis in order to extend to calculate excited states. In this section, a simpler, more general approach is developed that can handle the ground and excited states as well as the spatially degenerate states (2s, $2p_x$, $2p_z$) within a general heterostructure potential, that might also include an electric field (bias).

As in equation (5.4) the Hamiltonian of an impurity atom within a semiconductor heterostructure under the single-band effective mass and envelope function approximations is:

$$\mathcal{H} = -\frac{\hbar^2}{2}\nabla\frac{1}{m^*}\nabla + V(z) - \frac{e^2}{4\pi\epsilon r} \tag{5.137}$$

where, again, m^* is the effective mass of the charge carrier, $V(z)$ is the electrostatic potential which defines the heterostructure (and may include a bias $-eFz$) and r is the distance between the impurity and the charge carrier. Placing the x and y origins on the impurity

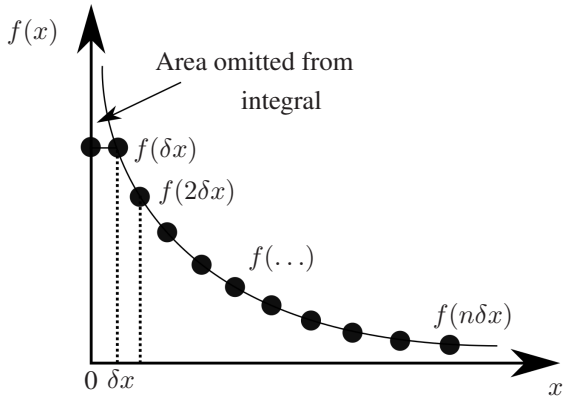


Figure 5.23: Samples (represented by closed circles) taken from a function $f(x)$, in which a singularity is present at $x = 0$. The sample at $x = 0$ has been replaced by $f(\delta x)$ to avoid the singularity. The resulting error in the area is highlighted

atom, which is at a position r_i, then:

$$r^2 = x^2 + y^2 + (z - r_i)^2 \tag{5.138}$$

Taking a trial wave function of a charge carrier (electron or hole):

$$\Psi = \psi(z)\xi(r) \tag{5.139}$$

where $\psi(z)$ is the wave function of the electron (or hole) in the same heterostructure but *without* the impurity present and $\xi(r)$ is a hydrogenic-like term describing the interaction between the electron (or hole) and the donor (or acceptor) ion.

For the 1s (ground) state of the impurity, the interaction term would be given by:

$$\xi(r) = \exp\left(-\frac{r}{\lambda}\right) \tag{5.140}$$

where λ is a variational parameter.

The variational calculation is implemented by adjusting λ in order to minimise the expectation value of the Hamiltonian operator (the total energy):

$$E = \frac{\langle\Psi|\mathcal{H}|\Psi\rangle}{\langle\Psi|\Psi\rangle} \tag{5.141}$$

The energy E is evaluated for different values of λ by direct numerical integration of the numerator and denominator in the above equation. For example:

$$\langle\Psi|\mathcal{H}|\Psi\rangle = \iiint_{-\infty}^{\infty} \Psi\left(-\frac{\hbar^2}{2}\nabla\frac{1}{m^*}\nabla + V(z) - \frac{e^2}{4\pi\epsilon r}\right)\Psi \,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \tag{5.142}$$

These integrals can be calculated using standard numerical methods (e.g. a simple strip summation) over a three-dimensional uniform mesh, with the differentials in the kinetic

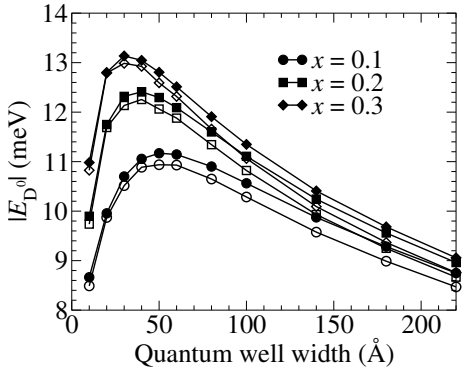


Figure 5.24: Comparison of the results of calculations using the approach developed in this section (open symbols) with the earlier method (closed symbols) presented in Fig. 5.11. The system is a donor at the centre of a GaAs quantum well surrounded by $Ga_{1-x}Al_xAs$ barriers

energy component evaluated using finite-difference expansions. However, the number of data points in the mesh may need to be relatively large in order to handle the infinite limits in each dimension, and this potentially leads to time-consuming calculations. Some improvement can be obtained by noting the symmetry of the Coulombic term and the hydrogenic wave functions in equation (5.142). Since the integrand is always even-symmetrical in both x and y, the integration domain may be reduced by half in each dimension to become:

$$\langle\Psi|\mathcal{H}|\Psi\rangle = 4\int_{-\infty}^{\infty}\mathrm{d}z\int_{0}^{\infty}\mathrm{d}x\int_{0}^{\infty}\mathrm{d}y\,\Psi\left(-\frac{\hbar^2}{2}\nabla\frac{1}{m^*}\nabla + V(z) - \frac{e^2}{4\pi\epsilon r}\right)\Psi \qquad (5.143)$$

This expression requires only a quarter of the number of data points to be computed, and as such is much faster to solve.

A complication arises, since the Coulomb potential $-e^2/(4\pi\epsilon r)$ has a singularity (i.e. an infinitely large value) at $r = 0$. It is, therefore, important to avoid sampling the integrand at this point, otherwise the entire numerical integration of equation (5.143) will evaluate to infinity. One simple technique for overcoming this problem is to approximate the integrand at $r = 0$ with its neighbouring sample (i.e. its value at $r = \delta x$). This has the effect of 'slicing the top' off the infinitely high function, as illustrated in Fig. 5.23, thus ensuring that a finite integral value will be found. Clearly, this results in an underestimate of the integral, but in practice this is a relatively small error if δx is narrow.

5.13 Direct evaluation of the expectation value of the Hamiltonian for the ground state

Figure 5.24 shows a comparison between the two numerical approaches of the ground state binding energy of a donor at the centre of a GaAs quantum well surrounded by $Ga_{1-x}Al_xAs$ barriers. The closed symbols represent the data calculated using the three-dimensional trial

wave function in Section 5.4 and presented in Fig. 5.11, while the open symbols show the results of calculations with the more direct numerical integration approach just developed.

It can be seen that for all three barrier heights (defined by three different Al concentrations, $x = 0.1$, 0.2 and 0.3), the function form of the binding energies with the quantum well width is very similar. However, in all cases, the new data are around $0.2\,\mathrm{meV}$ below the earlier calculations.

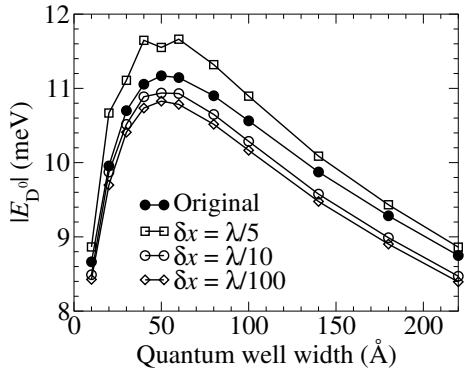


Figure 5.25: The effect of the numerical integration step length δx on the binding energy of a donor at the centre of a GaAs quantum well surrounded by $Ga_{0.9}Al_{0.1}As$ barriers

In Fig. 5.25, the in-plane integration step length δx (equal to δy) was taken as one-fifth, one-tenth and one-hundredth of the Bohr radius λ. The results of the calculations are compared to the $x = 0.1$ data from the original method. It can be seen that decreasing the size of the integration step length δx (which will increase the accuracy of any numerical integrations) does indeed result in the new data converging to a value around $0.3\,\mathrm{meV}$ lower than that of the earlier method. However, it also increases computation time, and given that this represents integration over a plane, increasing the number of steps by a factor of 10 (from $\lambda/10$ to $\lambda/100$, say), increases the computation time by a factor of 100. The step length $\delta x = \lambda/10$ is a good compromise and is adopted in all subsequent calculations.

5.14 Validation of the model for the position dependence of the impurity

It has thus been established that this approach to the calculation of the ground state energy level of an impurity at the centre of a quantum well is in good agreement with the earlier technique. However, it is now important to validate the new approach for various impurity positions—not restricting the impurity to the highly symmetric position at the centre of a quantum well. This is particularly important because of the choice of the impurity wave function as the product of the envelope function without the impurity present with

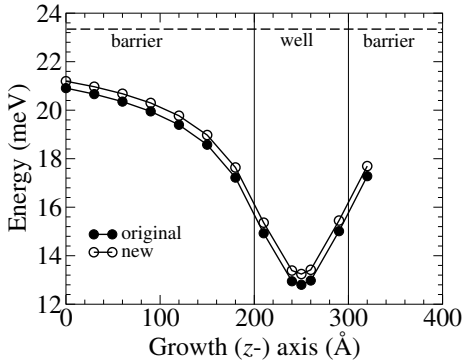


Figure 5.26: Comparison of the results of calculations using the approach developed in this section (open symbols) with the earlier method (closed symbols). The figure shows the total energy of an electron in a $100\,$Å GaAs quantum well surrounded by $200\,$Å $Ga_{0.9}Al_{0.1}As$ barriers, as a function of the position of the donor across the barrier–well–barrier system

the hydrogenic term, i.e. $\Psi = \psi(z)\xi(x, y, z, r_i)$, rather than the original method which recalculated the envelope function.

Figure 5.26 shows the results of a series of calculations of the ground state donor energy for a range of donor positions across the barrier, well and barrier regions of a single $100\,$Å GaAs quantum well surrounded by $200\,$Å barriers of $Ga_{0.9}Al_{0.1}As$. The figure also shows the results from the earlier method, and it can be seen that the difference in the energy of the donor is around $0.5\,\mathrm{meV}$, which is quite acceptable in device design.

The calculations span the quantum well to show that the expected symmetry of the energy with position is reproduced. The horizontal dashed line near the top of the figure shows the energy of an electron in the same heterostructure but without a donor present.

5.15 Excited states

Thus, it now remains to move onto the initial purpose of this work, which is to develop a model of the energy levels of *excited* impurity states. Taking the hydrogenic factor ξ of the total impurity wave function as:

$$\xi = \left(1 - \frac{r}{\lambda}\right) \exp\left(-\frac{r}{\lambda}\right) \tag{5.144}$$

where $r = \sqrt{x^2 + y^2 + (z - r_i)^2}$, the energy level of the excited 2s state of a donor in the centre of a GaAs quantum well can be calculated. Strictly speaking, the total impurity wave function $\Psi = \psi(z)\xi(x, y, z, r_i)$ for this excited 2s state is not orthogonal to the corresponding 1s wave function and a more complicated approach incorporating a factor α in ξ should be used. However, the results that follow indicate that the simple form taken in equation (5.144) is a good approximation for obtaining the *energy* of excited impurity states in heterostructures.

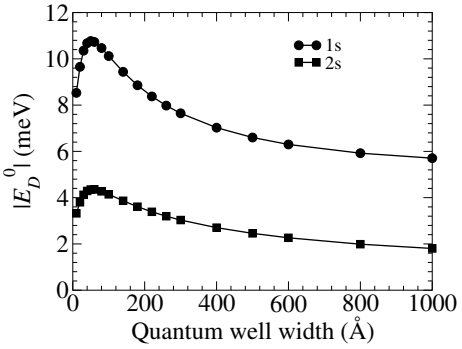


Figure 5.27: The well width dependence of the binding energies of the 1s and 2s states of donors at the centre of GaAs quantum wells surrounded by $Ga_{0.9}Al_{0.1}As$ barriers

Figure 5.27 shows the binding energy of the 1s and 2s states as a function of the width of the GaAs quantum well. It can be seen that the binding energy of the 1s state is much larger than that of the 2s (as expected) and that the functional form of both curves is the same. Figure 5.28 shows the ratio of these two binding energies. Bohr theory [2] gives the (binding) energy of the hydrogen atom as $E_n = E_1/n^2$, where E_1 is the (binding) energy of the lowest ($n = 1$) state. Thus for a hydrogen atom the ratio of the energy of the 1s state to the 2s state is exactly 4. In the figure, this number is much smaller; however, as the width of the well is increased, and the system tends towards bulk, this ratio does begin to move towards 4. This is supporting evidence in favour of the application of this method to the calculation of the binding energy of excited impurity states in heterostructures.

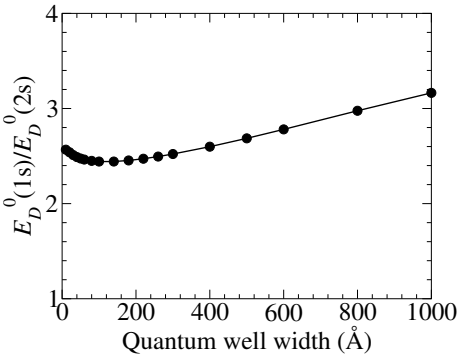


Figure 5.28: The well width dependence of the ratio of the binding energies of the 1s and 2s states of donors at the centre of GaAs quantum wells surrounded by $Ga_{0.9}Al_{0.1}As$ barriers

Figure 5.29 shows the binding energies of the 2s and $2p_x$ impurity levels for a donor at the centre of a GaAs quantum well, as a function of the quantum well width; again the barrier material is $Ga_{0.9}Al_{0.1}As$. The hydrogenic factor for the $2p_x$ state was taken as:

$$\xi = x \exp\left(-\frac{r}{\lambda}\right) \tag{5.145}$$

It can be seen that the binding energy of the $2p_x$ state is considerably less than that of the 2s state—only around half around the peak at 100 Å. However, as the width of the quantum well is increased, the binding energy of the $2p_x$ level converges towards that of the 2s level, as would be expected, because in the limit of an infinitely wide well (i.e. bulk material), these states should be degenerate.

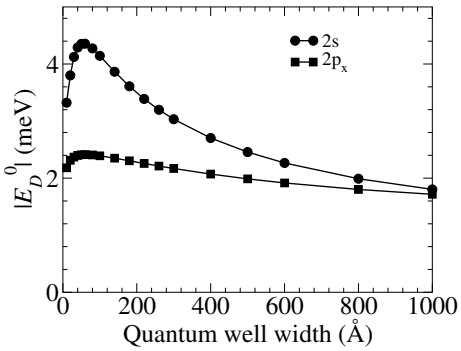


Figure 5.29: The well width dependence of the binding energies of the 2s and $2p_x$ states of donors at the centre of GaAs quantum wells surrounded by $Ga_{0.9}Al_{0.1}As$ barriers

Figure 5.30 shows the Bohr radii (the value of the variational parameter) λ for the 1s and $2p_x$ states. Figure 5.31 shows the ratio of these two radii. It can be seen that it is very close to 2—as expected from Bohr theory. In fact, the $2p_x$ state is a very good one to calculate because near the impurity, where the singularity exists, the functional form of the hydrogenic term, i.e. $x \exp(-r/\lambda)$, naturally tends to zero, hence reducing any inaccuracies that might arise because of integrating over the singularity.

Figure 5.32 shows the binding energy of the $2p_z$ impurity state, calculated using a hydrogenic term analogous to that for the $2p_x$ state, as in equation (5.145), but with the 'lobes' aligned along the growth axis. The lobes now overlap with the repulsive potentials of the heterostructure barriers and hence the energy of the impurity state is increased. This *decreases* the binding energy, and in fact, as the figure shows, the binding energy becomes negative, i.e. the state is unbound. This effect is reduced as the well becomes wider and the z-lobes can 'fit' inside the well. The binding energy of the $2p_z$ state tends to that of the 2s and $2p_x$ states as the well approaches 1000 Å.

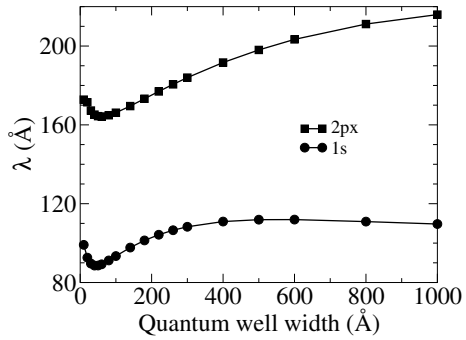


Figure 5.30: The well width dependence of the Bohr radius λ of the 1s and $2p_x$ states of donors at the centre of GaAs quantum wells surrounded by $Ga_{0.9}Al_{0.1}As$ barriers

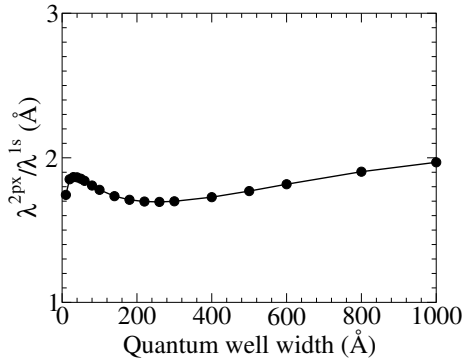


Figure 5.31: The well width dependence of the *ratio* of the Bohr radii of the 1s and $2p_x$ states of donors at the centre of GaAs quantum wells surrounded by $Ga_{0.9}Al_{0.1}As$ barriers

5.16 Impurity occupancy statistics

Any book on solid-state physics, e.g. Kasap [32] (p. 314) or Sze [5] (p. 16), gives the number of free electrons introduced into a crystal due to doping as the integral over the entire conduction band of the density of states multiplied by the probability of occupation:

$$n = \int_{E_c}^{\infty} \rho(E) f(E)\, \mathrm{d}E \tag{5.146}$$

where the density of states $\rho(E)$ (derived earlier in Section 2.4) is written in relation to the conduction-band edge E_c as:

$$\rho(E) = \frac{1}{2\pi^2} \left(\frac{2m^*}{\hbar^2} \right)^{\frac{3}{2}} (E - E_c)^{\frac{1}{2}} \tag{5.147}$$

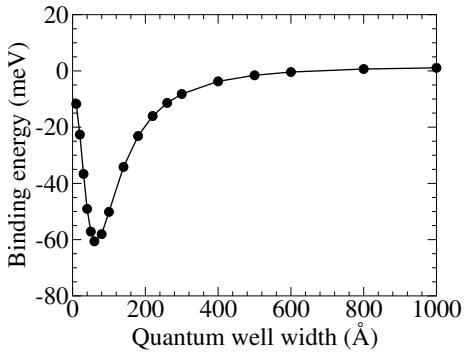


Figure 5.32: The well width dependence of the binding energy of the $2p_z$ state of donors at the centre of GaAs quantum wells surrounded by $Ga_{0.9}Al_{0.1}As$ barriers

and the Fermi–Dirac[3] distribution function $f(E)$ is approximated by a Boltzmann distribution function:

$$f(E) = \exp\left(-\frac{(E - E_{\mathrm{F}})}{kT}\right) \tag{5.148}$$

The number of free electrons (which is the same as the number of ionised donors) then follows as:

$$n = \frac{1}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{\frac{3}{2}} \int_{E_{\mathrm{c}}}^{\infty} (E - E_{\mathrm{c}})^{\frac{1}{2}} \exp\left(-\frac{(E - E_{\mathrm{F}})}{kT}\right) \mathrm{d}E \tag{5.149}$$

which leads to (see Kasap [32], p. 314):

$$n = N_{\mathrm{c}} \exp\left(-\frac{(E_{\mathrm{c}} - E_{\mathrm{F}})}{kT}\right) \tag{5.150}$$

where N_{c} is known as the 'effective density of states' and is given by:

$$N_{\mathrm{c}} = 2\left(\frac{2\pi m^* kT}{h^2}\right)^{\frac{3}{2}} \tag{5.151}$$

and the temperature-dependent Fermi energy for a donor volume density N_{d} and energy E_{d} is given by:

$$E_{\mathrm{F}} = \frac{E_{\mathrm{c}} + E_{\mathrm{d}}}{2} + \frac{1}{2}kT \ln\left(\frac{N_{\mathrm{d}}}{2N_{\mathrm{c}}}\right) \tag{5.152}$$

(see, for example, Kasap [32], p. 335). It can be seen from equation (5.152) that, at low temperatures, E_{F} would be halfway between the donor energy levels E_{d} and the conduction-band edge E_{c} as expected.

[3]This is the true Fermi–Dirac distribution function for a doped semiconductor and not the quasi-function introduced earlier to describe the distribution of electrons within a subband.

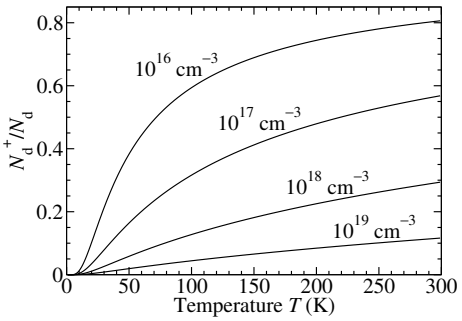


Figure 5.33: The proportion of ionised donors N_{d}^{+} from a doping concentration N_{d} as a function of temperature for a bulk semiconductor, for several different dopant densities. The material system is GaAs with a donor energy $5.8\,\mathrm{meV}$ below the conduction-band edge and an effective mass of $0.067m_0$

Figure 5.33 shows the results of calculations of the number of ionised donors N_{d}^{+} (=n) as a function of the temperature T for several different doping densities N_{d}. It can be seen from the figure that as the doping density increases the proportion of donors ionised decreases at any given temperature. This is just a result of statistical mechanics—the effective density of states N_{c} remains constant, so as more donors are added to the system a smaller *proportion* of them can find empty states for their electrons to move to. It is interesting to note that at the typical doping density of $1 \times 10^{18}\,\mathrm{cm}^{-3}$ only around one-third of donors are ionised at room temperature.

The focus of this book is on low-dimensional systems, so generalising to the two dimensions of a quantum well heterostructure, the density of states would become the two-dimensional density of states again derived in Section 2.4:

$$\rho^{\mathrm{2D}}(E) = \frac{m^*}{\pi\hbar^2} \tag{5.153}$$

Hence:

$$n = \frac{m^*}{\pi\hbar^2}\int_{E_{\mathrm{c}}}^{\infty} \exp\left(-\frac{(E-E_{\mathrm{F}})}{kT}\right)\mathrm{d}E = \frac{m^*}{\pi\hbar^2}\left[-kT\exp\left(-\frac{(E-E_{\mathrm{F}})}{kT}\right)\right]_{E_{\mathrm{c}}}^{\infty} \tag{5.154}$$

$$\therefore n = \frac{m^*kT}{\pi\hbar^2}\exp\left(-\frac{(E_{\mathrm{c}}-E_{\mathrm{F}})}{kT}\right) \tag{5.155}$$

which in analogy to three dimensions, could also be written:

$$n = N_{\mathrm{c}}^{\mathrm{2D}}\exp\left(-\frac{(E_{\mathrm{c}}-E_{\mathrm{F}})}{kT}\right) \tag{5.156}$$

where the two-dimensional effective density of states is given by:

$$N_{\mathrm{c}}^{\mathrm{2D}} = \frac{m^*kT}{\pi\hbar^2} \tag{5.157}$$

and the Fermi energy would be equivalent to that in equation (5.152), but with N_c^{2D} instead of N_c, and the donor density N_d would be taken as a *sheet density*.

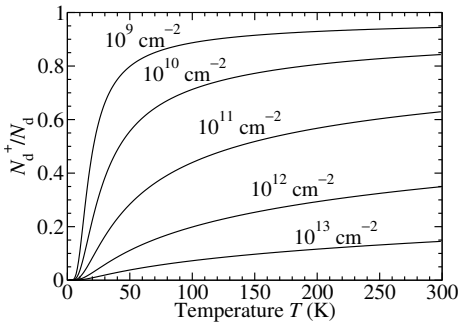


Figure 5.34: The proportion of ionised donors N_d^+ from a doping concentration N_d as a function of temperature for a two-dimensional (quantum well) system, for several different dopant densities. The material system is GaAs with a donor energy $5.8\,\mathrm{meV}$ below the conduction-band edge and an effective mass of $0.067m_0$

Figure 5.34 shows the corresponding curve to Fig. 5.33 for a two-dimensional system. The generic behaviour is the same as in bulk, i.e. the higher the temperature the greater the proportion of donors ionised, and the greater the doping density the lower the proportion of donors ionised. Again, the number of free carriers generated by doping might seem surprisingly small.

Although this view has been recently reinforced [33], it is actually too simplistic and the proportion of carriers ionised is too small. The reason is that the model assumes that the impurity ionisation energy (which in this section would be $E_c - E_d$) is left as a constant, whereas it was found that screening should see it reduce as the impurity concentration increases [34]. Pearson and Bardeen [34] argued that the ionisation energy of an impurity would be reduced by an energy inversely proportional to the distance between the impurities, hence the ionisation energy should be equal to:

$$|E_D^0| - aN_d^{\frac{1}{3}}$$

where E_D^0 is the ionisation (binding) energy of the isolated impurity and N_d is the impurity (donor) concentration. The constant a depends on the impurity species and the host and was deduced experimentally by Pearson and Bardeen for boron and phosphorus in silicon.

Exercises

(1). The Schrödinger equation for an electron orbiting a donor ion in a bulk semiconductor crystal of permittivity $\epsilon = \epsilon_{\rm r}\epsilon_0$ is:

$$-\frac{\hbar^2}{2m_e^*}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\psi(r)-\frac{e^2}{4\pi\epsilon r}\psi(r)=E\psi(r)$$

where the wave function is assumed to take the spherical symmetry of the potential and r is the distance of the electron from the positively charged donor ion (assumed to have infinite mass). By converting the kinetic energy operator into spherical polar coordinates and then substituting in a trial wave function $\psi(r) = \exp(-r/\lambda)$, find an expression for the Bohr radius λ of the electron orbit. Hence, derive the binding energy of a neutral donor as given in equation (5.1).

(2). When the concentration of donors in a bulk crystal increases to the point when the wave functions of the bound electrons start to overlap, transport through an impurity band can occur. This occurs when the average separation of the donors is of the order of the Bohr radius. Calculate this critical impurity density, sometimes known as the Mott transition, for GaAs.

(3). What would be the equivalent critical sheet (areal) density for impurity band transport in a single layer of donors as may be found in a GaAs quantum well?

(4). Confirm that the integral in equation (5.118) reduces to the 2D case (equation (5.59)) when $\zeta = 0$, and to the 3D case (equation (5.92)) when $\zeta = 1$. In the latter case, use the substitution variable $u = (w^2 + 1)/(2w)$ to solve the integral.

(5). Limit the Schrödinger equation in exercise (1) to two dimensions and, by following a similar procedure, calculate the binding energy of a two-dimensional neutral donor. Determine the ratio of the binding energies of the two-dimensional donor to the three-dimensional.

(6). Investigate how close the two-dimensional limit can be approached by placing a donor in the centre of an infinitely deep GaAs quantum well and compressing the electron wave function by decreasing the width of the well without limit. Compare the results using the variational method program (`qwwad_ef_donor_specific`) with the two-dimensional trial wave function in Section 5.3 (using the `--symmetry 2D` program option), the three-dimensional trial wave function of Section 5.4 (`--symmetry 3D`) and direct numerical integration of Section 5.12 (using the `qwwad_ef_donor_general` program).

(7). The method derived for computing the impurity binding energy within a quantum well assumes the permittivity $\epsilon = \epsilon_0\epsilon_{\rm r}$ is constant. Make a simple estimation of the maximum error this could introduce for a donor at the centre of a GaAs quantum well surrounded by $Ga_{0.8}Al_{0.2}As$ barriers.

(8). Use the computer program `qwwad_ef_donor_general` to calculate the 2s, $2p_x$ and $2p_z$ donor binding-energy states in a GaAs/$Ga_{0.9}Al_{0.1}As$ quantum well as a function of the well width, as in Section 5.15. However, go beyond the range of large quantum well widths used in the text in order to pay particular attention to the convergence, or

otherwise, of these different excited donor state energies and Bohr radii as the quantum well width increases.

References

[1] S. M. Sze, *Semiconductor Devices: Physics and Technology*, Wiley, New York, 1985.

[2] R. T. Weidner and R. L. Sells, *Elementary Modern Physics*, Allyn and Bacon, Boston, Third edition, 1980.

[3] L. I. Schiff, *Quantum Mechanics*, McGraw-Hill, New York, 1968.

[4] R. M. Eisberg, *Fundamentals of Modern Physics*, John Wiley & Sons, Inc., New York, 1961.

[5] S. M. Sze, *Physics of Semiconductor Devices*, John Wiley & Sons, Inc., New York, Second edition, 1981.

[6] S. Adachi, *GaAs and Related Materials: Bulk Semiconducting and Superlattice Properties*, World Scientific, Singapore, 1994.

[7] P. Harrison, F. Long, and W. E. Hagston, 'Empirical pseudo-potential calculation of the in-plane effective masses of electron and holes of two-dimensional excitons in CdTe quantum wells', *Superlatt. Microstruct.*, **19**:123–130, 1996.

[8] J. S. Blakemore, *Solid State Physics*, Cambridge University Press, Cambridge, Second edition, 1985.

[9] N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Saunders College Publishing, Philadelphia, 1976.

[10] C. Mailhiot, Y.-C. Chang, and T. C. McGill, 'Energy spectra of donors in GaAs/$Ga_{1-x}Al_x$ As quantum well structures in the effective mass approximation', *Phys. Rev. B*, **26**:4449, 1982.

[11] R. L. Greene and K. K. Bajaj, 'Energy levels of hydrogenic impurity states in GaAs-$Ga_{1-x}As_x$As quantum well structures', *Solid State Commun.*, **45**:825, 1983.

[12] G. N. Carneiro, G. Weber, and L. E. Oliveira, 'Binding energies and intra-donor absorption spectra in GaAs–GaAlAs quantum wells', *Semicond. Sci. Technol.*, **10**(1):41, 1995.

[13] S. Chaudhuri and K. K. Bajaj, 'Effect of non-parabolicity on the energy levels of hydrogenic donors in GaAs–$Ga_{1-x}Al_x$As quantum-well structures', *Phys. Rev. B*, **29**(4):1803–1806, 1984.

[14] C.-R. Huo, B.-Y. Gu, and L. Gu, 'General variational expressions for the calculation of the binding energies of anisotropic donor states in stepped quantum wells', *J. Appl. Phys.*, **70**(8):4357–4361, 1991.

[15] R. G. Roberts, P. Harrison, and W. E. Hagston, 'The symmetry of donor–bound electron wavefunctions in quantum wells', *Superlatt. Microstruct.*, **23**(2):289–296, 1998.

[16] W. E. Hagston, P. Harrison, and T. Stirner, 'Neutral donors and spin-flip Raman spectra in dilute magnetic semiconductor microstructures', *Phys. Rev. B*, **49**:8242–8248, 1994.

[17] G. Bastard, *Wave Mechanics Applied to Semiconductor Heterostructures*, Monographies de physique. Halsted Press, New York, 1988.

[18] W. T. Masselink, Y.-C. Chang, and H. Morkoç, 'Binding-energies of acceptors in GaAs–Al_xGa_{1-x}As quantum-wells', *Phys. Rev. B*, **28**(12):7373–7376, 1983.

[19] R. C. Miller, A. C. Gossard, W. T. Tsang, and O. Munteanu, 'Extrinsic photoluminescence from GaAs quantum wells', *Phys. Rev. B*, **25**:3871, 1982.

[20] S. Fraizzoli, F. Bassani, and R. Buczko, 'Shallow donor impurities in GaAs-$Ga_{1-x}Al_x$As quantum well structures: Role of the dielectric constant mismatch', *Phys. Rev. B*, **41**:5096, 1990.

[21] U. Ekenberg, 'Nonparabolicity effects in a quantum well: Sublevel shift, parallel mass, and Landau levels', *Phys. Rev. B*, **40**(11):7714–7726, 1989.

[22] S. R. Parihar and S. A. Lyon, 'Optical transitions and energy level ordering for quantum confined impurities,' in *Quantum Well Intersubband Transition Physics and Devices*, H. C. Liu, B. F. Levine, and J. Y. Andersson, Eds., vol. 270 of *NATO ASI*, pp. 403–410. Springer Netherlands, Dordrecht, 1994.

[23] J. K. Furdyna, J. Kossut, R. K. Willardson, and A. C. Beer, Eds., *Diluted Magnetic Semiconductors*, Number 25 in Semiconductors and Semimetals. Academic Press, Boston, 1988.

[24] A. Twardowski, *Diluted Magnetic Semiconductors*, World Scientific, Singapore, 1996.

[25] J. M. Fatah, T. Piorek, P. Harrison, T. Stirner, and W. E. Hagston, 'Numerical simulation of anti-ferromagnetic spin-pairing effects in diluted magnetic semiconductors and enhanced paramagnetism at interfaces', *Phys. Rev. B*, **49**:10341–10344, 1994.

[26] J. A. Gaj, C. Bodin-Deshayes, P. Peyla, G. F. J. Cibert, Y. M. d'Aubigne, R. Romestain, and A. Wasiela, 'Magneto-optical study of interface mixing in CdTe-(Cd,Mn)Te system', in *Proceedings of the 21st International Conference on the Physics of Semiconductors*, Singapore, 1992, p. 1936, World Scientific.

[27] A. K. Ramdas and S. Rodriguez, 'Effect of strain on vibrational modes in strained layer superlattices', in *Semiconductors and Semimetals*, R. K. Willardson and A. C. Beer, Eds., Boston, 1988, vol. 25, p. 345, Academic Press.

[28] M. P. Halsall, S. V. Railson, D. Wolverson, J. J. Davies, B. Lunn, and D. E. Ashenford, 'Spin-flip Raman scattering in $CdTe/Cd_{1-x}Mn_xTe$ multiple quantum wells: A model system for the study of electron-donor binding in semiconductor heterostructures', *Phys. Rev. B*, **50**(16):11755–11763, 1994.

[29] D. R. Yakovlev, 'Two dimensional magnetic polarons in semimagnetic quantum well structures', in *Festkörperprobleme/Advances in Solid State Physics*, U. Roessler, Ed., Braunschweig, 1992, vol. 32, p. 251, Vieweg.

[30] W. E. Hagston, P. Harrison, J. H. C. Hogg, S. Jackson, J. E. Nicholls, T. Stirner, B. Lunn, and D. E. Ashenford, 'Molecular-beam epitaxy investigation of interface disorder effects in magnetic II–VI quantum wells', *J. Vac. Sci. Technol. B*, **11**(3):881–884, 1993.

[31] T. Stirner, P. Harrison, W. E. Hagston, and J. P. Goodwin, 'Band gap renormalization and observation of the type I–type II transition in quantum well systems', *J. Appl. Phys.*, **73**(10):5081–5087, 1993.

[32] S. O. Kasap, *Principles of Electrical Engineering Materials and Devices*, Irwin McGraw-Hill, Boston, 1997.

[33] G. Xiao, J. Lee, J. J. Liou, and A. Ortiz-Conde, 'Incomplete ionization in a semiconductor and its implications to device modeling', *Microelectronics Reliability*, **39**:1299, 1999.

[34] G. L. Pearson and J. Bardeen, 'Electrical properties of pure silicon and silicon alloys containing boron and phosphorous', *Phys. Rev.*, **75**:865, 1949.

6

Excitons

6.1 Excitons in bulk

If photons of energy comparable to the band gap are incident on a semiconductor, then they can be absorbed by the electrons forming atomic bonds between neighbouring atoms, providing them with enough energy to break free and move around in the body of the crystal. Within the band theory of solids, this would be described as 'exciting an electron from the valence band across the band gap into the conduction band'. If the energy of the photon is larger than the band gap, then a free electron is created *and* an empty state is left within the valence band (see 'high-energy excitation' in Fig. 6.1). The empty state within the valence band behaves very much like an air bubble in a liquid and rises to the top—the lowest-energy state. This 'hole' behaves as though it were positively charged and hence often forms a bond with a conduction-band electron (see 'exciton formation' in Fig. 6.1). The attractive potential leads to a reduction (by an amount E_{X^0}) in the total energy of the electron and hole. This bound electron–hole pair is known as an *exciton*. Photons of energy just below the band gap can be absorbed, thus creating excitons directly (see 'resonant excitation' in Fig. 6.1).

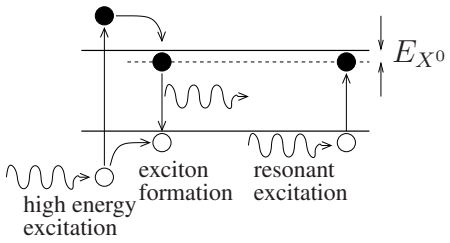


Figure 6.1: Schematic representation of the non-resonant and resonant generation of excitons

As the hole mass is generally much greater than the electron mass, the two-body system resembles a hydrogen atom, with the negatively charged electron orbiting the positive hole. The exciton is quite stable and can have a relatively long lifetime, of the order of hundreds of picoseconds to nanoseconds. Exciton recombination is an important feature of low-temperature photoluminescence, although as the binding energies are relatively low, i.e. a few millielectronvolts to a few tens of millielectronvolts, they tend to dissociate at higher temperatures.

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

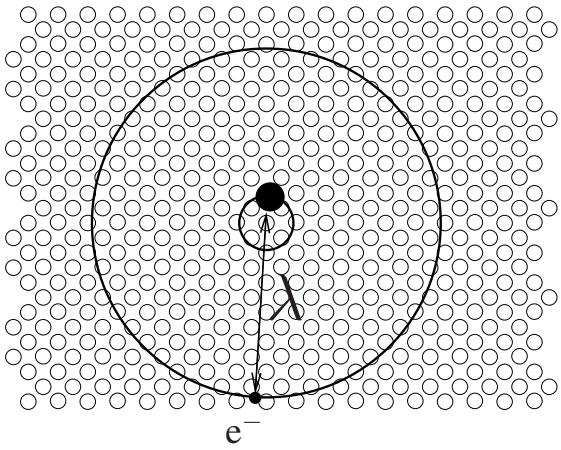


Figure 6.2: Schematic diagram of an exciton in bulk material, with the hole (filled circle near centre) and electron 'separated' by a Bohr radius λ, orbiting around the centre of mass

Therefore, in a similar manner to the hydrogenic impurities discussed in Chapter 5, the binding energy and orbital radius can be represented well by Bohr theory, *with* the correction for the finite mass of the central charge [1, 2]. This is implemented by exchanging the (in the case of a hydrogen atom, or donor) orbiting electron mass with the reduced mass of the two-body system, in this case the electron–hole pair. The reduced mass is given by:

$$\frac{1}{\mu} = \frac{1}{m_{\mathrm{e}}^*} + \frac{1}{m_{\mathrm{h}}^*} \tag{6.1}$$

Hence, the binding energy becomes:

$$E_{\mathrm{X}^0} = -\frac{\mu e^4}{32\pi^2\hbar^2\epsilon_{\mathrm{r}}^2\epsilon_0^2} \tag{6.2}$$

and the Bohr radius follows as:

$$\lambda = \frac{4\pi\epsilon_{\mathrm{r}}\epsilon_0\hbar^2}{\mu e^2} \tag{6.3}$$

Taking typical values for bulk GaAs, i.e. the Γ valley electron and heavy-hole effective masses, $m_{\mathrm{e}}^* = 0.067m_0$ and $m_{\mathrm{hh}}^* = 0.62m_0$, respectively, then $\mu = 0.060m_0$. Using the static dielectric constant $\epsilon_{\mathrm{r}} = 13.18$ [3], then the exciton binding energy and Bohr radius respectively follow as:

$$E_{\mathrm{X}^0} = -4.7\,\mathrm{meV} \quad \text{and} \quad \lambda = 115\,\text{Å} \tag{6.4}$$

which are exactly the same as the low-temperature measured values (see [3], p. 420). Correspondingly for CdTe, $m_{\mathrm{e}}^* = 0.096m_0$ and $m_{\mathrm{hh}}^* = 0.6m_0$, therefore $\mu = 0.083m_0$, with $\epsilon_{\mathrm{r}} = 10.6$, then:

$$E_{\mathrm{X}^0} = -10.1\,\mathrm{meV} \quad \text{and} \quad \lambda = 67\,\text{Å} \tag{6.5}$$

which again agree with experiment [4].

6.2 Excitons in heterostructures

In the same way as in bulk, excitons can be formed by the bonding of free electron–free hole pairs or through resonant excitation. Whereas in bulk the total energy of the exciton is simply the energy of the free electron–free hole pair (i.e. the band gap) plus the exciton binding energy E_{X^0}, in a heterostructure there are additional components due to the electron and hole confinement energies, i.e.

$$E = E_{\mathrm{g}} + E_{\mathrm{X}^0} \text{ (bulk)}, \qquad E = E_{\mathrm{g}} + E_{\mathrm{e}} + E_{\mathrm{h}} + E_{\mathrm{X}^0} \text{ (heterostructure)} \tag{6.6}$$

The total exciton energy is clearly a function of structure because of the structural

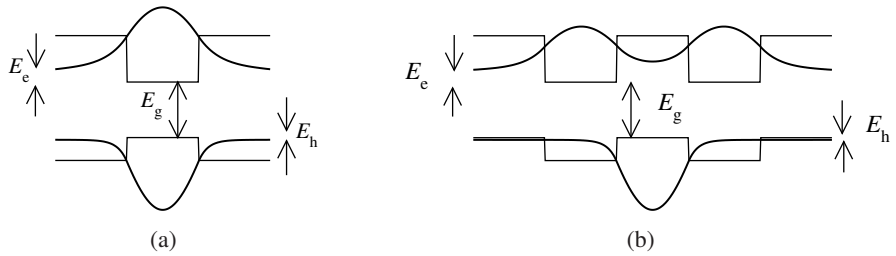


Figure 6.3: Electrons and holes in (a) a Type-I single quantum well and (b) a Type-II multiple quantum well system

dependency of the confinement energies. In addition, it must be expected that the Coulombic potential energy E_{X^0} also depends upon the structure. The latter effect arises because the electron–hole separation can vary quite considerably between heterostructures. Figure 6.3 illustrates this schematically; clearly the electron–hole separation is much smaller in the Type-I quantum well, where both particles are localised in the same layer of semiconductor, than in the Type-II system, where they are localised in different layers. Hence, the exciton binding energy will be larger in the former.

6.3 Exciton binding energies

The Hamiltonian representing the interacting two-body electron–hole complex can be considered to be the sum of three terms

$$\mathcal{H} = \mathcal{H}_{\mathrm{e}} + \mathcal{H}_{\mathrm{h}} + \mathcal{H}_{\mathrm{e\text{–}h}} \tag{6.7}$$

where $\mathcal{H}_{\mathrm{e}}$ and $\mathcal{H}_{\mathrm{h}}$ are the one-particle Hamiltonians appropriate to the conduction and valence bands, respectively, of the particular microstructure of interest (as in Chapter 2). However, in the context of the exciton a constant effective mass along the growth (z-)axis must be assumed, i.e. the one-particle Hamiltonians are written:

$$\mathcal{H}_{\mathrm{e}} = -\frac{\hbar^2}{2m_{\mathrm{e}}^*}\frac{\partial^2}{\partial z^2} + V_{\mathrm{CB}}(z), \qquad \mathcal{H}_{\mathrm{h}} = -\frac{\hbar^2}{2m_{\mathrm{h}}^*}\frac{\partial^2}{\partial z^2} + V_{\mathrm{VB}}(z) \tag{6.8}$$

The third term, $\mathcal{H}_{\text{e–h}}$, represents the electron–hole interaction, and is itself composed of two terms. One of these corresponds to the kinetic energy of the relative motion of the electron and hole in the x–y plane (perpendicular to the growth axis), while the other represents the Coulombic potential energy, i.e.

$$\mathcal{H}_{\text{e–h}} = \frac{\mathcal{P}_\perp^2}{2\mu_\perp} - \frac{e^2}{4\pi\epsilon r} \tag{6.9}$$

where $\mathcal{P}_\perp$ is the quantum mechanical momentum operator for the in-plane component of the relative motion. Now:

$$\mathcal{P} = -i\hbar\nabla = -i\hbar\left(\frac{\partial}{\partial x}\hat{\mathbf{i}} + \frac{\partial}{\partial y}\hat{\mathbf{j}} + \frac{\partial}{\partial z}\hat{\mathbf{k}}\right) \tag{6.10}$$

$$\therefore \mathcal{P}_\perp^2 = -\hbar^2\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) \tag{6.11}$$

and r is simply the electron–hole separation, given by:

$$r^2 = (x_{\text{e}} - x_{\text{h}})^2 + (y_{\text{e}} - y_{\text{h}})^2 + (z_{\text{e}} - z_{\text{h}})^2 \tag{6.12}$$

Note, that as the main motivation is concerned with the optical properties of excitons in quantum well systems, the Hamiltonian contains no term for the motion of the centre of mass of the exciton in the x–y plane, i.e. the exciton is assumed to be at rest within the plane of the well.

The problem, as always, is to find the eigenfunctions Ψ and eigenvalues E of the Schrödinger equation:

$$\mathcal{H}\Psi = E\Psi \tag{6.13}$$

which describes the system. Following standard procedures, the two-body exciton wave function Ψ is chosen to be a product of three factors:

$$\Psi = \psi_{\text{e}}(z_{\text{e}})\psi_{\text{h}}(z_{\text{h}})\psi_{\text{r}} \tag{6.14}$$

where ψ_{r} represents the electron and hole relative motion. Following Hilton *et al*. [5, 6] and Harrison *et al*. [7, 8], ψ_{r} will be a variational wave function employed to minimise the total energy E of the system. The specific form of ψ_{r} will be discussed later. The other two factors, $\psi_{\text{e}}(z_{\text{e}})$ and $\psi_{\text{h}}(z_{\text{h}})$, are simply the eigenfunctions of the one-particle Hamiltonians of the heterostructure:

$$\mathcal{H}_{\text{e}}\psi_{\text{e}}(z_{\text{e}}) = E_{\text{e}}\psi_{\text{e}}(z_{\text{e}}), \qquad \mathcal{H}_{\text{h}}\psi_{\text{h}}(z_{\text{h}}) = E_{\text{h}}\psi_{\text{h}}(z_{\text{h}}) \tag{6.15}$$

One of the main advantages of this formalism is that it is independent of the form of the one-particle Hamiltonians $\mathcal{H}_{\text{e}}$ and $\mathcal{H}_{\text{h}}$, and indeed calculations can be performed on any system in which the standard electron and hole wave functions can be calculated [9, 10, 11].

Multiplying equation (6.13) on the left by Ψ and integrating over all space, the total exciton energy follows simply as the expectation value:

$$E = \frac{\langle\Psi|\mathcal{H}|\Psi\rangle}{\langle\Psi|\Psi\rangle} \tag{6.16}$$

Now, whatever the form of the relative motion term ψ_{r} of equation (6.14), it will be a function of the electron–hole separation, which is quantified in terms of the three variables, $x = |x_{\mathrm{e}} - x_{\mathrm{h}}|$, $y = |y_{\mathrm{e}} - y_{\mathrm{h}}|$ and $a = |z_{\mathrm{e}} - z_{\mathrm{h}}|$. The denominator $\mathcal{D}$ in equation (6.16) can therefore be written as:

$$\mathcal{D} = \langle \Psi | \Psi \rangle = \int_{\text{all space}} |\psi_{\mathrm{e}}(z_{\mathrm{e}})|^2 |\psi_{\mathrm{h}}(z_{\mathrm{h}})|^2 |\psi_{\mathrm{r}}|^2 \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z_{\mathrm{e}} \, \mathrm{d}z_{\mathrm{h}} \tag{6.17}$$

Since ψ_{e} and ψ_{h} are functions of z only, the integrations over the x–y plane will involve only the relative motion term ψ_{r}. Correspondingly, the integrations over the electron and hole coordinates, z_{e} and z_{h}, will yield a result which is a function of the electron and hole separation (a) only, i.e. it is possible to write:

$$\int |\psi_{\mathrm{r}}|^2 \, \mathrm{d}x \, \mathrm{d}y = F(a) \tag{6.18}$$

Although the one-particle wave functions, ψ_{e} and ψ_{h}, in principle, extend to $\pm\infty$, in practice, from a computational viewpoint, it is necessary to define them between 'effective infinities'. As discussed in Chapter 3, these are defined so that:

$$\int_{\mu}^{\nu} \psi_{\mathrm{e}}(z_{\mathrm{e}}) \, \mathrm{d}z_{\mathrm{e}} \approx \int_{-\infty}^{+\infty} \psi_{\mathrm{e}}(z_{\mathrm{e}}) \, \mathrm{d}z_{\mathrm{e}} \tag{6.19}$$

and similarly for ψ_{h}. Using these results, the denominator $\mathcal{D}$ in equation (6.17) can be re-written as:

$$\begin{aligned} \mathcal{D} = \int_{0}^{\nu-\mu} \int_{\mu}^{\nu} \int_{\mu}^{\nu} |\psi_{\mathrm{e}}(z_{\mathrm{e}})|^2 |\psi_{\mathrm{h}}(z_{\mathrm{h}})|^2 F(a) \, \mathrm{d}z_{\mathrm{e}} \, \mathrm{d}z_{\mathrm{h}} \\ \times \left[\delta\left(z_{\mathrm{e}} - z_{\mathrm{h}} - a\right) + \delta\left(z_{\mathrm{h}} - z_{\mathrm{e}} - a\right) \right] \, \mathrm{d}a \end{aligned} \tag{6.20}$$

where the Dirac δ-functions have been introduced in order to ensure that contributions are included from both $z_{\mathrm{e}} - z_{\mathrm{h}} = a$ and $z_{\mathrm{h}} - z_{\mathrm{e}} = a$. If the integration over z_{h} is performed first, then the first term only has a finite value when $z_{\mathrm{h}} = z_{\mathrm{e}} - a$ and the second term when $z_{\mathrm{h}} = z_{\mathrm{e}} + a$. Hence:

$$\begin{aligned} \mathcal{D} = \int_{0}^{\nu-\mu} \int_{\mu+a}^{\nu} |\psi_{\mathrm{e}}(z_{\mathrm{e}})|^2 |\psi_{\mathrm{h}}(z_{\mathrm{e}} - a)|^2 F(a) \, \mathrm{d}z_{\mathrm{e}} \, \mathrm{d}a \\ + \int_{0}^{\nu-\mu} \int_{\mu}^{\nu-a} |\psi_{\mathrm{e}}(z_{\mathrm{e}})|^2 |\psi_{\mathrm{h}}(z_{\mathrm{e}} + a)|^2 F(a) \, \mathrm{d}z_{\mathrm{e}} \, \mathrm{d}a \end{aligned} \tag{6.21}$$

Making the substitution $z = z_{\mathrm{e}} - a$ in the first term, and relabelling z as z_{e} in the second, finally gives the denominator in equation (6.16) as:

$$\mathcal{D} = \int_{0}^{\nu-\mu} p(a) F(a) \, \mathrm{d}a \tag{6.22}$$

where $p(a)$ represents the uncorrelated probability of finding the electron and hole separated by a distance a, i.e.

$$p(a) = \int_{\mu}^{\nu-a} |\psi_{\mathrm{e}}(z+a)|^2 |\psi_{\mathrm{h}}(z)|^2 + |\psi_{\mathrm{e}}(z)|^2 |\psi_{\mathrm{h}}(z+a)|^2 \, \mathrm{d}z \tag{6.23}$$

Returning to the expression for the expectation value of the total exciton energy in equation (6.16) and using the three-term Hamiltonian of equation (6.7), the numerator $\mathcal{N}$ can then be written as:

$$\mathcal{N} = \langle\Psi|\mathcal{H}|\Psi\rangle = \langle\Psi|\mathcal{H}_{\mathrm{e}}|\Psi\rangle + \langle\Psi|\mathcal{H}_{\mathrm{h}}|\Psi\rangle + \langle\Psi|\mathcal{H}_{\mathrm{e\text{–}h}}|\Psi\rangle \tag{6.24}$$

Labelling these three terms as $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$, respectively, i.e. $\mathcal{N} = \mathcal{A} + \mathcal{B} + \mathcal{C}$, consider:

$$\mathcal{A} = \langle\Psi|\mathcal{T}|\Psi\rangle + \langle\Psi|\mathcal{V}|\Psi\rangle \tag{6.25}$$

where $\mathcal{T}$ and $\mathcal{V}$ are respectively the kinetic and potential energy operators of the electron one-particle Hamiltonian $\mathcal{H}_{\mathrm{e}}$. Employing the standard constant-mass kinetic operator leads (in an obvious notation) to:

$$\mathcal{A}_{\mathcal{T}} = \int \psi^*_{\mathrm{e}}(z_{\mathrm{e}})\psi^*_{\mathrm{h}}(z_{\mathrm{h}})\psi^*_{\mathrm{r}} \frac{-\hbar^2}{2m^*_{\mathrm{e}}} \frac{\partial^2}{\partial z^2_{\mathrm{e}}} \psi_{\mathrm{e}}(z_{\mathrm{e}})\psi_{\mathrm{h}}(z_{\mathrm{h}})\psi_{\mathrm{r}} \,\mathrm{d}z_{\mathrm{e}}\,\mathrm{d}z_{\mathrm{h}}\;\,\mathrm{d}x\,\mathrm{d}y \tag{6.26}$$

On performing the differentiations:

$$\begin{aligned}\mathcal{A}_{\mathcal{T}} = \int \psi^*_{\mathrm{e}}(z_{\mathrm{e}})\psi^*_{\mathrm{h}}(z_{\mathrm{h}})\psi^*_{\mathrm{r}} \frac{-\hbar^2}{2m^*_{\mathrm{e}}} \big[\psi''_{\mathrm{e}}(z_{\mathrm{e}})\psi_{\mathrm{h}}(z_{\mathrm{h}})\psi_{\mathrm{r}} + 2\psi'_{\mathrm{e}}(z_{\mathrm{e}})\psi_{\mathrm{h}}(z_{\mathrm{h}})\psi'_{\mathrm{r}} \\ + \psi_{\mathrm{e}}(z_{\mathrm{e}})\psi_{\mathrm{h}}(z_{\mathrm{h}})\psi''_{\mathrm{r}}\big]\,\mathrm{d}z_{\mathrm{e}}\,\mathrm{d}z_{\mathrm{h}}\,\mathrm{d}x\,\mathrm{d}y\end{aligned} \tag{6.27}$$

which can be written as follows:

$$\begin{aligned}\mathcal{A}_{\mathcal{T}} = \int \frac{-\hbar^2}{2m^*_{\mathrm{e}}} \big[\psi^*_{\mathrm{e}}(z_{\mathrm{e}})\psi''_{\mathrm{e}}(z_{\mathrm{e}})|\psi_{\mathrm{h}}(z_{\mathrm{h}})|^2|\psi_{\mathrm{r}}|^2 \\ + 2\psi^*_{\mathrm{e}}(z_{\mathrm{e}})\psi'_{\mathrm{e}}(z_{\mathrm{e}})|\psi_{\mathrm{h}}(z_{\mathrm{h}})|^2\psi^*_{\mathrm{r}}\psi'_{\mathrm{r}} + |\psi_{\mathrm{e}}(z_{\mathrm{e}})|^2|\psi_{\mathrm{h}}(z_{\mathrm{h}})|^2\psi^*_{\mathrm{r}}\psi''_{\mathrm{r}}\big]\,\mathrm{d}z_{\mathrm{e}}\,\mathrm{d}z_{\mathrm{h}}\,\mathrm{d}x\,\mathrm{d}y\end{aligned} \tag{6.28}$$

Note that the first term, when normalised by the denominator $\mathcal{D}$, merely represents the one-particle kinetic energy of the electron, which coupled together with $\langle\Psi|\mathcal{V}|\Psi\rangle$ in equation (6.25) gives the one-particle electron energy E_{e} as defined in equation (6.15). Furthermore, noting that, for the problems of interest, i.e. stationary states within semiconductor microstructures, the wave functions are real and that the chain rule gives:

$$\psi_{\mathrm{e}}(z_{\mathrm{e}})\psi'_{\mathrm{e}}(z_{\mathrm{e}}) = \frac{1}{2}\frac{\partial\psi^2_{\mathrm{e}}(z_{\mathrm{e}})}{\partial z_{\mathrm{e}}} \tag{6.29}$$

equation (6.25) becomes:

$$\begin{aligned}\mathcal{A} = E_{\mathrm{e}}\mathcal{D} + \int \frac{-\hbar^2}{2m^*_{\mathrm{e}}} \Bigg\{|\psi_{\mathrm{h}}(z_{\mathrm{h}})|^2 \left[\frac{\partial\psi^2_{\mathrm{e}}(z_{\mathrm{e}})}{\partial z_{\mathrm{e}}}\right]\psi_{\mathrm{r}}\psi'_{\mathrm{r}} \\ + |\psi_{\mathrm{e}}(z_{\mathrm{e}})|^2|\psi_{\mathrm{h}}(z_{\mathrm{h}})|^2\psi_{\mathrm{r}}\psi''_{\mathrm{r}}\Bigg\}\,\mathrm{d}z_{\mathrm{e}}\,\mathrm{d}z_{\mathrm{h}}\,\mathrm{d}x\,\mathrm{d}y\end{aligned} \tag{6.30}$$

Integrating by parts over $\mathrm{d}z_{\mathrm{e}}$ the first term in the integrand, i.e.

$$\begin{aligned}\int |\psi_{\mathrm{h}}(z_{\mathrm{h}})|^2 \left[\frac{\partial|\psi_{\mathrm{e}}(z_{\mathrm{e}})|^2}{\partial z_{\mathrm{e}}}\right]\psi_{\mathrm{r}}\psi'_{\mathrm{r}}\,\mathrm{d}z_{\mathrm{e}} = |\psi_{\mathrm{h}}(z_{\mathrm{h}})|^2\left[|\psi_{\mathrm{e}}(z_{\mathrm{e}})|^2\psi_{\mathrm{r}}\psi'_{\mathrm{r}}\right]^{\nu}_{\mu} \\ - \int \left(\psi'^2_{\mathrm{r}} + \psi_{\mathrm{r}}\psi''_{\mathrm{r}}\right)|\psi_{\mathrm{e}}(z_{\mathrm{e}})|^2|\psi_{\mathrm{h}}(z_{\mathrm{h}})|^2\,\mathrm{d}z_{\mathrm{e}}\end{aligned} \tag{6.31}$$

Hence, substituting back into equation (6.30) gives:

$$\mathcal{A} = E_{\mathrm{e}}\mathcal{D} + \int \frac{-\hbar^2}{2m_{\mathrm{e}}^*}|\psi_{\mathrm{h}}(z_{\mathrm{h}})|^2 \Big\{ \left[|\psi_{\mathrm{e}}(z_{\mathrm{e}})|^2\psi_{\mathrm{r}}\psi_{\mathrm{r}}'\right]_{\mu}^{\nu}$$

$$- \int \left[|\psi_{\mathrm{e}}(z_{\mathrm{e}})|^2\psi_{\mathrm{r}}'^2 + |\psi_{\mathrm{e}}(z_{\mathrm{e}})|^2\psi_{\mathrm{r}}\psi_{\mathrm{r}}'' - |\psi_{\mathrm{e}}(z_{\mathrm{e}})|^2\psi_{\mathrm{r}}\psi_{\mathrm{r}}''\right] \ \mathrm{d}z_{\mathrm{e}} \ \Big\} \mathrm{d}z_{\mathrm{h}}\ \mathrm{d}x\ \mathrm{d}y \tag{6.32}$$

The last two terms in the integration over z_{e} obviously cancel out, and, in addition, examination of the first term on the right-hand side shows that this term will not contribute, since whatever the form of the eigenfunction ψ_{e}, it will undoubtedly vanish at the effective infinities μ and ν. Therefore:

$$\mathcal{A} = E_{\mathrm{e}}\mathcal{D} + \int_{\text{all space}} \frac{\hbar^2}{2m_{\mathrm{e}}^*}|\psi_{\mathrm{h}}(z_{\mathrm{h}})|^2|\psi_{\mathrm{e}}(z_{\mathrm{e}})|^2\psi_{\mathrm{r}}'^2\ \mathrm{d}z_{\mathrm{e}}\ \mathrm{d}z_{\mathrm{h}}\ \mathrm{d}x\ \mathrm{d}y \tag{6.33}$$

Adopting a similar argument as above, the integral over the plane can then be written as:

$$G(a) = \int \left|\frac{\partial\psi_{\mathrm{r}}}{\partial z_{\mathrm{e}}}\right|^2 \ \mathrm{d}x\ \mathrm{d}y \tag{6.34}$$

and hence equation (6.33) gives the final expression for $\mathcal{A}$ as follows:

$$\mathcal{A} = E_{\mathrm{e}}\mathcal{D} + \frac{\hbar^2}{2m_{\mathrm{e}}^*}\int_0^{\nu-\mu} p(a)G(a)\ \mathrm{d}a \tag{6.35}$$

An analogous expression exists for $\mathcal{B}$ with the subscripts e and h, labelling electron and hole respectively, being interchanged.

Now consider $\mathcal{C} = \langle\Psi|\mathcal{H}_{\text{e–h}}|\Psi\rangle$ as defined in equation (6.24). As mentioned earlier, $\mathcal{H}_{\text{e–h}}$ is composed of two terms which represent the in-plane kinetic energy of the relative motion and the Coulombic potential energy between the electron and hole, i.e.

$$\mathcal{C} = \mathcal{C}_{\mathcal{T}} + \mathcal{C}_{\mathcal{V}} \tag{6.36}$$

By using the kinetic energy operator defined above in equation (6.11), the first of these terms can then be written as:

$$\mathcal{C}_{\mathcal{T}} = \int \psi_{\mathrm{e}}^*(z_{\mathrm{e}})\psi_{\mathrm{h}}^*(z_{\mathrm{h}})\psi_{\mathrm{r}}^* \frac{-\hbar^2}{2\mu_{\perp}}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\psi_{\mathrm{e}}(z_{\mathrm{e}})\psi_{\mathrm{h}}(z_{\mathrm{h}})\psi_{\mathrm{r}}\ \mathrm{d}z_{\mathrm{e}}\ \mathrm{d}z_{\mathrm{h}}\ \mathrm{d}x\ \mathrm{d}y \tag{6.37}$$

where the in-plane reduced mass $\mu_{\perp}$ is given by:

$$\frac{1}{\mu_{\perp}} = \frac{1}{m_{\mathrm{e}}^{\perp}} + \frac{1}{m_{\mathrm{h}}^{\perp}} \tag{6.38}$$

Again, whatever the functional form of the relative motion term ψ_{r}, it is apparent that the kinetic energy operator in $\mathcal{C}_{\mathcal{T}}$ will act only upon ψ_{r}, as the one-particle electron and hole wave functions, ψ_{e} and ψ_{h} respectively, are not functions of x or y. Furthermore, the

integration over the plane will result in an entity that is a function of a only, i.e. in the spirit of equation (6.35):

$$\mathcal{C}_{\mathcal{T}} = -\frac{\hbar^2}{2\mu_\perp}\int_0^{\nu-\mu} p(a)J(a)\,\mathrm{d}a \tag{6.39}$$

where

$$J(a) = \int \psi_{\mathrm{r}}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\psi_{\mathrm{r}}\,\mathrm{d}x\,\mathrm{d}y \tag{6.40}$$

since ψ_{r} is real. In a similar manner, the potential energy term $\mathcal{C}_{\mathcal{V}}$ can be written as:

$$\mathcal{C}_{\mathcal{V}} = -\frac{e^2}{4\pi\epsilon}\int_0^{\nu-\mu} p(a)K(a)\,\mathrm{d}a \tag{6.41}$$

where

$$K(a) = \int \frac{1}{r}\psi_{\mathrm{r}}^2\,\mathrm{d}x\,\mathrm{d}y \tag{6.42}$$

with $r^2 = x^2 + y^2 + a^2$.

To summarise, then, the total exciton energy can be written as:

$$E = \frac{\mathcal{A} + \mathcal{B} + \mathcal{C}}{\mathcal{D}} \tag{6.43}$$

with $\mathcal{A}$ given in equation (6.35), $\mathcal{B}$ following analogously, $\mathcal{C}$ given by equation (6.36) and $\mathcal{D}$ by equation (6.22). Evaluation of the entities $F(a)$, $G(a)$, $J(a)$ and $K(a)$ necessitates the introduction of a specific form for the relative motion term ψ_{r}. The exciton binding energy $E_{\mathrm{X^0}}$ follows simply from:

$$E = E_{\mathrm{e}} + E_{\mathrm{h}} + E_{\mathrm{X^0}} \tag{6.44}$$

6.4 1s exciton

In line with other work in the literature, the choice of wave function representing the electron–hole interaction that is to be employed is a hydrogenic type, given by:

$$\psi_{\mathrm{r}} = \exp\left(-\frac{r'}{\lambda}\right) \tag{6.45}$$

where the Bohr radius λ will be used as a parameter and systematically varied in order to minimise the total energy E of the system (which is equivalent to maximising the exciton binding energy $E_{\mathrm{X^0}}$). However, where this work differs is in the precise form of the choice of the relative coordinate r'. Using the knowledge and experience derived from the consideration of impurities in Chapter 5, a variable symmetry-type relative motion term is chosen:

$$r'^2 = (x_{\mathrm{e}} - x_{\mathrm{h}})^2 + (y_{\mathrm{e}} - y_{\mathrm{h}})^2 + \zeta^2\,(z_{\mathrm{e}} - z_{\mathrm{h}})^2 \tag{6.46}$$

or, by using $a = |z_{\mathrm{e}} - z_{\mathrm{h}}|$ as above:

$$r'^2 = r_\perp^2 + \zeta^2 a^2 \tag{6.47}$$

The second variational parameter, ζ, allows the exciton to assume any shape of wave function that is energetically favourable. Traditionally, the case with $\zeta = 0$ has become known as the two-dimensional exciton [12, 13, 14, 15, 16], and $\zeta = 1$ as the three-dimensional exciton [12, 17]. Cases where ζ is allowed to take values other than 0 and 1 are rarely found in the literature because of the increased complexity in the resulting mathematics of the problem [18, 19, 20, 21].

In the present approach, ζ will be allowed to take all values from 0 upwards; however, due to complications in the mathematics, in particular deriving the limits of integration for the evaluation of $K(a)$ (see later), it is necessary to introduce the following transformations:

$$\text{for} \quad 0 \leq \zeta \leq 1, \quad \zeta^2 = 1 - \beta^2, \quad \text{where} \quad 0 \leq \beta \leq 1 \tag{6.48}$$

$$\text{for} \quad 1 \leq \zeta \leq \infty, \quad \zeta^2 = 1 + \eta^2, \quad \text{where} \quad 0 \leq \eta \leq \infty \tag{6.49}$$

In order to illustrate the technique, attention will be focused initially on the former case.

First, consider evaluation of the entity $F(a)$ as defined in equation (6.18). Using this particular choice of ψ_{r} and moving into plane polar coordinates gives:

$$F(a) = 2\pi \int_0^\infty \exp\left(-\frac{2r'}{\lambda}\right) \; r_\perp \, \mathrm{d}r_\perp \tag{6.50}$$

Using equation (6.47) to substitute $r_\perp$ with r', noting that $r'\mathrm{d}r' = r_\perp \mathrm{d}r_\perp$, and accounting for the change in limits, i.e. when:

$$r_\perp = 0, \quad r' = \sqrt{1-\beta^2}a \tag{6.51}$$

$$r_\perp = \infty, \quad r' = \infty \tag{6.52}$$

yields:

$$F(a) = 2\pi \int_{\sqrt{1-\beta^2}a}^\infty \exp\left(-\frac{2r'}{\lambda}\right) \; r' \, \mathrm{d}r' \tag{6.53}$$

and by integrating by parts:

$$F(a) = 2\pi \left(\frac{\lambda\sqrt{1-\beta^2}a}{2} + \frac{\lambda^2}{4}\right) \exp\left(-\frac{2\sqrt{1-\beta^2}a}{\lambda}\right) \tag{6.54}$$

Next consider $G(a)$ as defined in equation (6.34). Then:

$$G(a) = 2\pi \int_0^\infty \left| \frac{\partial}{\partial z_{\mathrm{e}}} \exp\left(-\frac{r'}{\lambda}\right) \right|^2 \; r_\perp \, \mathrm{d}r_\perp \tag{6.55}$$

Performing the differentiation and substituting r' for $r_\perp$ as above leads to:

$$G(a) = 2\pi \int_{\sqrt{1-\beta^2}a}^\infty \left| -\frac{1}{\lambda} \exp\left(-\frac{r'}{\lambda}\right) \frac{(1-\beta^2)a}{r'} \right|^2 \; r' \, \mathrm{d}r' \tag{6.56}$$

Now let $r' = \sqrt{1-\beta^2}a\cosh\theta$. Then:

$$G(a) = 2\pi \int_0^\infty \frac{1}{\lambda^2}(1-\beta^2)^2 a^2 \exp\left(-\frac{2\sqrt{1-\beta^2}a\cosh\theta}{\lambda}\right) \frac{\sqrt{1-\beta^2}a\sinh\theta}{\sqrt{1-\beta^2}a\cosh\theta} \, \mathrm{d}\theta \tag{6.57}$$

Making the further substitution $w = \exp(-\theta)$, so that $\mathrm{d}\theta = -\mathrm{d}w/w$, and noting that:

$$\cosh\theta = \frac{1}{2}\left(e^{\theta} + e^{-\theta}\right) = \frac{1}{2}\left(\frac{1}{w} + w\right), \qquad \sinh\theta = \frac{1}{2}\left(\frac{1}{w} - w\right) \tag{6.58}$$

then finally:

$$G(a) = 2\pi \int_0^1 \frac{(1-\beta^2)^2 a^2}{\lambda^2} \exp\left[-\frac{\sqrt{1-\beta^2}a}{\lambda}\left(\frac{1}{w} + w\right)\right]\left(\frac{1-w^2}{w(1+w^2)}\right)\,\mathrm{d}w \tag{6.59}$$

An analogous argument shows that $G(a)$ appearing in $\mathcal{B}$ is exactly equal to this form.

Next consider evaluation of the integral $J(a)$ as defined in equation (6.40). With this aim, note that:

$$\frac{\partial r'}{\partial x} = \frac{x}{r'}, \qquad \frac{\partial^2 r'}{\partial x^2} = \frac{1}{r'} - \frac{x^2}{r'^3} \tag{6.60}$$

This then gives:

$$\frac{\partial^2 \psi_{\mathrm{r}}}{\partial x^2} = \left(\frac{-1}{r'\lambda} + \frac{x^2}{r'^3\lambda} + \frac{x^2}{r'^2\lambda^2}\right)\psi_{\mathrm{r}} \tag{6.61}$$

and hence:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\psi_{\mathrm{r}} = \left[\frac{1}{\lambda^2} - \frac{(1-\beta^2)a^2}{\lambda r'^3} - \frac{1}{r'\lambda} - \frac{(1-\beta^2)a^2}{r'^2\lambda^2}\right]\psi_{\mathrm{r}} \tag{6.62}$$

Therefore, moving from Cartesian into plane polar coordinates:

$$J(a) = 2\pi \int_0^{\infty} \left[\frac{1}{\lambda^2} - \frac{(1-\beta^2)a^2}{\lambda r'^3} - \frac{1}{r'\lambda} - \frac{(1-\beta^2)a^2}{r'^2\lambda^2}\right] \exp\left(-\frac{2r'}{\lambda}\right)\, r_{\perp}\,\mathrm{d}r_{\perp} \tag{6.63}$$

It is standard practice in the literature [22] to expand the expressions involving $(r')^{-n}$ as a power series in $r_{\perp}$ and a, and then to perform the integration numerically. This involves summing over a series of terms, each of which must be integrated over a range from 0 to ∞. This procedure can, however, be avoided.

Writing equation (6.63) as $J(a) = J_1 + J_2 + J_3 + J_4$, where J_i represent the first, second, ... terms, then:

$$J_1 + J_3 = 2\pi \int_0^{\infty} \left[\frac{1}{\lambda^2} - \frac{1}{\lambda r'}\right] \exp\left(-\frac{2r'}{\lambda}\right)\, r_{\perp}\,\mathrm{d}r_{\perp} \tag{6.64}$$

Again, substituting r' for $r_{\perp}$:

$$J_1 + J_3 = 2\pi \int_{\sqrt{1-\beta^2}a}^{\infty} \left[\frac{1}{\lambda^2} - \frac{1}{\lambda r'}\right] \exp\left(-\frac{2r'}{\lambda}\right)\, r'\,\mathrm{d}r' \tag{6.65}$$

and therefore:

$$J_1 + J_3 = 2\pi \left\{\left[\frac{1}{2}\exp\left(-\frac{2r'}{\lambda}\right) - \frac{r'}{2\lambda}\exp\left(-\frac{2r'}{\lambda}\right)\right]_{\sqrt{1-\beta^2}a}^{\infty}\right.$$

$$+\frac{1}{2\lambda}\int_{\sqrt{1-\beta^2}a}^{\infty}\exp\left(-\frac{2r'}{\lambda}\right)\,\mathrm{d}r'\Bigg\} \tag{6.66}$$

which gives:

$$J_1+J_3=2\pi\left(\frac{\sqrt{1-\beta^2}a}{2\lambda}-\frac{1}{4}\right)\exp\left(-\frac{2\sqrt{1-\beta^2}a}{\lambda}\right) \tag{6.67}$$

The two remaining terms of equation (6.63) give:

$$J_2+J_4=2\pi\int_0^{\infty}\left[-\frac{(1-\beta^2)a^2}{\lambda r'^3}-\frac{(1-\beta^2)a^2}{\lambda^2 r'^2}\right]\,\exp\left(-\frac{2r'}{\lambda}\right)\,r_\perp\,\mathrm{d}r_\perp \tag{6.68}$$

Again substituting r' for $r_\perp$, then:

$$J_2+J_4=2\pi\int_{\sqrt{1-\beta^2}a}^{\infty}(1-\beta^2)a^2\left(-\frac{1}{\lambda r'^2}-\frac{1}{\lambda^2 r'}\right)\exp\left(-\frac{2r'}{\lambda}\right)\,\mathrm{d}r' \tag{6.69}$$

Making a further scale change of $r'=\sqrt{1-\beta^2}a\cosh\theta$:

$$J_2+J_4=2\pi\int_0^{\infty}\left(-\frac{1}{\lambda\cosh^2\theta}-\frac{\sqrt{1-\beta^2}a}{\lambda^2\cosh\theta}\right)\exp\left(-\frac{2\sqrt{1-\beta^2}a\cosh\theta}{\lambda}\right)$$
$$\times\sqrt{1-\beta^2}a\sinh\theta\,\mathrm{d}\theta \tag{6.70}$$

With the final substitution of $w=\exp(-\theta)$, with $\mathrm{d}\theta=-\mathrm{d}w/w$, therefore:

$$J_2+J_4=2\pi\int_0^1\left[-\frac{1}{\frac{\lambda}{4}\left(\frac{1}{w}+w\right)^2}-\frac{\sqrt{1-\beta^2}a}{\frac{\lambda^2}{2}\left(\frac{1}{w}+w\right)}\right]$$
$$\times\exp\left[-\frac{\sqrt{1-\beta^2}a}{\lambda}\left(\frac{1}{w}+w\right)\right]\sqrt{1-\beta^2}\frac{a}{2}\left(\frac{1}{w^2}-1\right)\,\mathrm{d}w \tag{6.71}$$

This last equation illustrates the advantage of the present formalism, namely that the computationally difficult integral of equation (6.63), which has hitherto been expanded into an infinite series and integrated to infinity, has been replaced with a simple integral over the range from 0 to 1. Even if the integrand had a finite number of singularities, this would still pose no problem in its evaluation.

Finally, consider $K(a)$, as defined in equation (6.42), i.e.

$$K(a)=2\pi\int_0^{\infty}\frac{1}{r}\exp\left(-\frac{2r'}{\lambda}\right)\,r_\perp\,\mathrm{d}r_\perp \tag{6.72}$$

Recalling that $r^2=r_\perp^2+a^2$, then:

$$K(a)=2\pi\int_a^{\infty}\exp\left(-\frac{2r'}{\lambda}\right)\,\mathrm{d}r \tag{6.73}$$

The form of r', i.e. $r'^2 = r^2 - \beta^2 a^2$, suggests the substitution $r = \beta a \cosh\theta$, which gives:

$$K(a) = 2\pi \int_{\cosh^{-1}\frac{1}{\beta}}^{\infty} \beta a \sinh\theta \exp\left(-\frac{2\beta a \sinh\theta}{\lambda}\right) \mathrm{d}\theta \tag{6.74}$$

Again making use of the substitution $w = \exp(-\theta)$, this then necessitates evaluating w corresponding to $\theta = \cosh^{-1}\frac{1}{\beta}$, i.e. $\cosh\theta = \frac{1}{\beta} = \frac{1}{2}(w + \frac{1}{w})$. This yields the quadratic equation:

$$w^2 - \frac{2w}{\beta} + 1 = 0 \tag{6.75}$$

Since the product of the two roots of this equation is unity, one root must correspond to $\exp(-\theta)$ and the other to $\exp\theta$. It is readily ascertained that the $\exp(-\theta)$ root is as follows:

$$\exp(-\theta) = \frac{1}{\beta} - \frac{1}{\beta}\sqrt{1-\beta^2} \tag{6.76}$$

This follows since the limits on β are 0 and 1. Hence:

$$K(a) = 2\pi\frac{\beta a}{2} \int_0^{\frac{1}{\beta}-\frac{1}{\beta}\sqrt{1-\beta^2}} \exp\left[-\frac{\beta a}{\lambda}\left(\frac{1}{w} - w\right)\right]\left(\frac{1}{w^2} - 1\right) \mathrm{d}w \tag{6.77}$$

In a similar manner, the second form of ψ_{r}, with $r' = \sqrt{r_\perp^2 + (1+\eta^2)a^2}$, gives the same expressions for $F(a)$, $G(a)$ and $J(a)$ as above, but with the simple substitution, $1 + \eta^2$, in place of $1 - \beta^2$. Only the expression for the last of the a-functions differs, in particular:

$$K(a) = 2\pi\frac{\eta a}{2} \int_0^{\frac{1}{\eta}\sqrt{1+\eta^2}-\frac{1}{\eta}} \exp\left[-\frac{\eta a}{\lambda}\left(\frac{1}{w} + w\right)\right]\left(\frac{1}{w^2} + 1\right) \mathrm{d}w \tag{6.78}$$

6.5 The two-dimensional and three-dimensional limits

It is *always* worthwhile performing convergence tests, i.e. taking the theoretical-computational model to established, often analytical, limits. The idea is to increase confidence in the theory and, as ever, to demonstrate that the previous theories are limits of the new. For example, classical mechanics is recovered from relativistic mechanics, in the low-velocity limit.

Although not as grand an example, there exist two limits which the analysis above can be compared with. In the limit of very wide quantum wells, the exciton should look like a bulk exciton, in both its binding energy and Bohr radius. In addition, in the limit of very narrow wells, the exciton should become two-dimensional in nature.

The bulk or three-dimensional limiting case of hydrogenic two-body systems, such as impurities and excitons, has been discussed and used already. A transparent treatise on this Bohr model of the hydrogen atom is given by Weidner and Sells [1]. This approach can easily be adapted to the case of an electron orbiting a positively charged central infinite mass, with the orbit restricted to a single plane, i.e. it is two-dimensional. The Schrödinger equation is then written as:

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2}\right) - \frac{e^2}{4\pi\epsilon_{\mathrm{r}}\epsilon_0 r_\perp}\psi = E\psi \tag{6.79}$$

where the electron–proton separation is:

$$r_\perp = \sqrt{x^2 + y^2} \tag{6.80}$$

For solution, the Cartesian coordinates need to be converted into the plane polar coordinate $r_\perp$. Note that:

$$\frac{\partial \psi}{\partial x} = \frac{\partial \psi}{\partial r_\perp}\frac{\partial r_\perp}{\partial x} = \frac{x}{r_\perp}\frac{\partial \psi}{\partial r_\perp} \tag{6.81}$$

and so:

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{x}{r_\perp}\frac{\partial \psi}{\partial r_\perp}\right) = \frac{1}{r_\perp}\frac{\partial \psi}{\partial r_\perp} - \frac{x}{r_\perp^2}\frac{\partial r_\perp}{\partial x}\frac{\partial \psi}{\partial r_\perp} + \frac{x}{r_\perp}\frac{\partial^2 \psi}{\partial r_\perp^2}\frac{\partial r_\perp}{\partial x} \tag{6.82}$$

$$\therefore \frac{\partial^2 \psi}{\partial x^2} = \frac{1}{r_\perp}\frac{\partial \psi}{\partial r_\perp} - \frac{x^2}{r_\perp^3}\frac{\partial \psi}{\partial r_\perp} + \frac{x^2}{r_\perp^2}\frac{\partial^2 \psi}{\partial r_\perp^2} \tag{6.83}$$

Thus:

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} = \frac{2}{r_\perp}\frac{\partial \psi}{\partial r_\perp} - \frac{x^2 + y^2}{r_\perp^3}\frac{\partial \psi}{\partial r_\perp} + \frac{x^2 + y^2}{r_\perp^2}\frac{\partial^2 \psi}{\partial r_\perp^2} \tag{6.84}$$

$$\therefore \frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} = \frac{1}{r_\perp}\frac{\partial \psi}{\partial r_\perp} + \frac{\partial^2 \psi}{\partial r_\perp^2} \tag{6.85}$$

Therefore the Schrödinger equation 6.79), in plane polar coordinates, becomes:

$$\frac{1}{r_\perp}\frac{\partial \psi}{\partial r_\perp} + \frac{\partial^2 \psi}{\partial r_\perp^2} + \frac{2m}{\hbar^2}\left(E + \frac{e^2}{4\pi\epsilon_\mathrm{r}\epsilon_0 r_\perp}\right)\psi = 0 \tag{6.86}$$

The standard technique for solving the hydrogen atom is to choose a spherically symmetric wave function of the form:

$$\psi = \exp\left(-\frac{r}{\lambda_{3\mathrm{D}}}\right) \tag{6.87}$$

In analogy to this, the 2D equivalent is:

$$\psi = \exp\left(-\frac{r_\perp}{\lambda_{2\mathrm{D}}}\right) \tag{6.88}$$

Calculating the derivatives:

$$\frac{\partial \psi}{\partial r_\perp} = -\frac{1}{\lambda_{2\mathrm{D}}}\psi \quad \text{and} \quad \frac{\partial^2 \psi}{\partial r_\perp^2} = \frac{1}{\lambda_{2\mathrm{D}}^2}\psi \tag{6.89}$$

and substituting into equation (6.86) yields:

$$\left(\frac{1}{\lambda_{2\mathrm{D}}^2} + \frac{2mE}{\hbar^2}\right)\psi + \frac{1}{r_\perp}\left(\frac{2me^2}{\hbar^2 4\pi\epsilon_\mathrm{r}\epsilon_0} - \frac{1}{\lambda}\right)\psi = 0 \tag{6.90}$$

The standard argument is that as this equation must be valid for all values of $r_\perp$, divergence as $r_\perp \to 0$ can only be avoided if the second term is zero. This implies:

$$\lambda_{2\mathrm{D}} = \frac{4\pi\epsilon_\mathrm{r}\epsilon_0\hbar^2}{2me^2} = \frac{\lambda_{3\mathrm{D}}}{2} \tag{6.91}$$

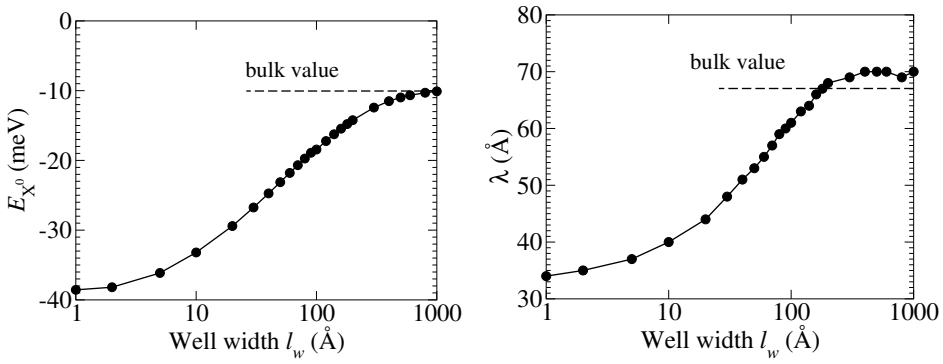


Figure 6.4: (a) Exciton binding energy and (b) Bohr radius in an infinitely deep quantum well

i.e. the Bohr radius of the 2D (planar) two-body system is half that of the 3D (spherical) system. Using this form for the Bohr radius, the first term then yields the energy of the 2D system as follows:

$$E^{\mathrm{2D}} = -4\frac{me^4}{32\pi^2\hbar^2\epsilon_{\mathrm{r}}^2\epsilon_0^2} = 4E^{\mathrm{3D}} \tag{6.92}$$

As mentioned above, and as already utilised for neutral impurities in Chapter 5, the 3D (or bulk) limit can be approached by increasing the width of a *finite* quantum well. However, the finite depth of such a quantum well is not conducive to the 2D limit, for as the well width decreases, the one-particle electron and hole wave functions begin to 'spill out' over the top of the well. The 2D limit can only be approached hypothetically in an infinitely deep quantum well. Figure 6.4(a) illustrates the results of calculations of the exciton binding energy as a function of the width of an infinitely deep CdTe quantum well. The magnitude of the bulk exciton binding energy was calculated in Section 6.1 as $10.1\,\mathrm{meV}$. The negative values on the graph illustrate that it is indeed a bound state.

Clearly, both limits are obeyed, i.e.

$$\lim_{l_w\to\infty} E_{\mathrm{X}^0} = E_{\mathrm{X}^0}^{\mathrm{3D}} \quad \text{and} \quad \lim_{l_w\to 0} E_{\mathrm{X}^0} = 4E_{\mathrm{X}^0}^{\mathrm{3D}} \tag{6.93}$$

Figure 6.4(b) displays the corresponding Bohr radii λ for the energies of Fig. 6.4(a). Remembering that the Bohr radius in bulk $\lambda_{\mathrm{3D}} = 67\,\text{Å}$, then the 2D limit, i.e.

$$\lim_{l_w\to 0} \lambda = \frac{\lambda_{\mathrm{3D}}}{2} \tag{6.94}$$

is satisfied. The 3D limit *is* obeyed, although the data on the graph show a slight scatter around the bulk radii of $67\,\text{Å}$. The source of this discrepancy is numerical accuracy. At the larger well widths, the wave function needs to be known at many points in order to calculate the binding energy to very high tolerances (thus leading to long computation times), as required here in these convergence tests. Despite this high accuracy, the Bohr radius still

shows some deviation from the bulk, although greater numerical accuracy has reduced this towards the value of 67 Å.

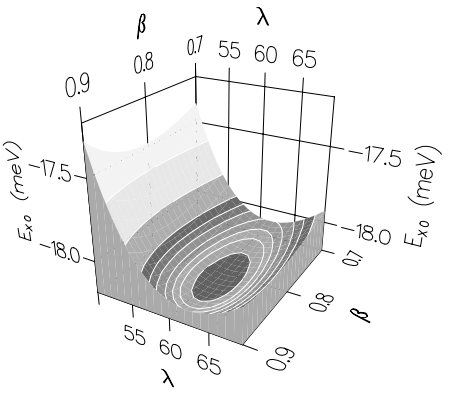


Figure 6.5: The exciton binding energy E_{X^0} in the two-dimensional λ–β parameter space

The calculations also showed that at the very narrow (5–40 Å) well widths, the second parameter ζ, representing the symmetry of the relative motion term, was close to zero, i.e. a 2D wave function. As the width increased, ζ increased steadily towards unity, i.e. the 3D (spherical or bulk-like) exciton. This point, that ζ generally lies in the limits $0 < \zeta < 1$, will be reinforced further by the use of several different example systems in the following discussions. For this reason, discussion of the symmetry of the wave function will use ζ and $\beta = \sqrt{1-\zeta^2}$ interchangeably.

Figure 6.5 illustrates the variational technique in the case of the infinite well of width l_w=100 Å. It can be seen that there is just one minimum in the exciton energy in the two-dimensional λ–β space. This is almost always the case. However, there is an exception to this rule [8]. In double quantum well systems where the one-particle electron wave function has a significant component in each of the wells, there exist two local energy minima within λ–β space. This implies that there are two possible exciton states which can be formed from the same one-particle wave functions. The two states can be thought of as originating from the localised hole binding separately with both of the 'lobes' of the electron wave function.

It is worthwhile just looking at the form of the function $p(a)$, which plays a significant role in the theoretical analysis. As mentioned at the point of its introduction, $p(a)$ represents the *uncorrelated* (i.e. calculated without taking account of the electron–hole Coulombic interaction) probability of finding the electron and hole separated by a distance a along the growth (z-)axis. While $p(a)$ is not a physical observable, often its form can help in understanding and provide an insight into the nature of the exciton. One particular example of that would be in the 'twin' exciton states of the double quantum well discussed above. Figure 6.6 shows $p(a)$ for this simple case of the infinite quantum well, while later figures will compare it with other more complex heterostructures.

In conclusion, the validity of the technique presented above has been substantiated by the calculation of exciton binding energies and wave functions for the full width range of infinite

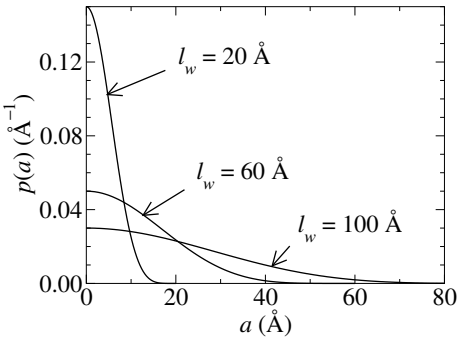


Figure 6.6: The uncorrelated probability $p(a)$ of finding the electron and hole separated by a distance a in the infinite quantum well, shown for different well widths

quantum wells. In the sections that follow, examples of calculations will be presented that demonstrate the versatility of the method. In addition, some of the results are expected to be of interest in their own right.

6.6 Excitons in single quantum wells

Only in a very few circumstances are the band offsets in both the conduction and valence bands large enough that the system can be approximated with an infinite quantum well model. The vast majority of cases demand finite potentials. Reiterating, the approach developed above depends only upon the one-particle electron and hole wave functions, $\psi_{\rm e}$ and $\psi_{\rm h}$, respectively, and no further knowledge of the system is required (except for the basic material parameters of electron and hole effective masses and dielectric constants).

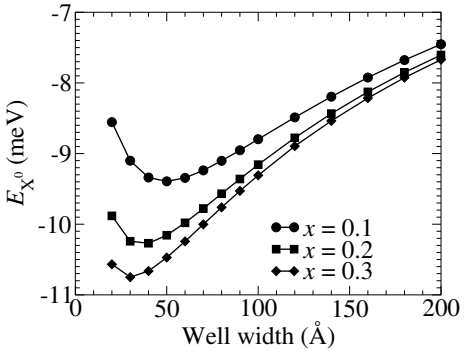


Figure 6.7: The exciton binding energy $E_{\rm X^0}$ in a GaAs quantum well surrounded by $\rm Ga_{1-x}Al_xAs$ barriers

To illustrate this, Figs. 6.7–6.9 display the results of calculations of exciton binding energies in finite GaAs single quantum wells surrounded by $Ga_{1-x}Al_xAs$ barriers. Unlike the infinite well case, the exciton binding energy E_{X^0} is a non-monotonic function of well width. As in the impurity binding energy case of Chapter 5, this is due to the effect of the well width on the electron and hole confinement. At very narrow well widths, the one-particle states are 'squeezed' up the well to reside at energies just below the top of the barrier. The wave function tends to 'spill' over the top, thus leading to a reduced probability of the particle being within the region of the quantum well. This non-monotonic behaviour has been observed in experiment [23]. In addition, increasing Al concentration in the barrier (both the conduction- and valence-band offsets are proportional to x) leads to an increase in the magnitude of the exciton binding energy for *all* well widths. However, there are two points to note regarding this. First, the effect of the barrier height is reduced at larger well widths. This substantiates the fact that a particular choice of barrier height (infinity) was chosen to illustrate the 3D limit in the previous section. It is clear from Fig. 6.7 that the exciton binding energy is tending towards its bulk value of $4.7\,\mathrm{meV}$ for *all* barrier heights. Secondly, the effect of an increasing barrier height is largest at smaller well widths, and for any given well width, further increases in the barrier height lead to smaller increases in the binding energy.

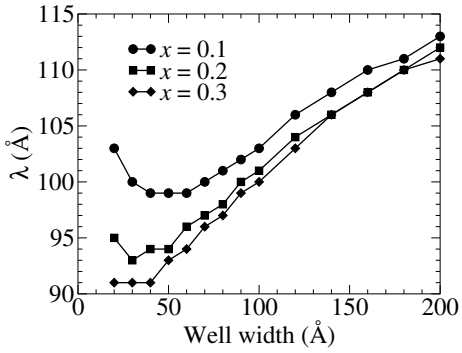


Figure 6.8: The exciton Bohr radius λ in a GaAs quantum well surrounded by $Ga_{1-x}Al_xAs$ barriers

Figures 6.8 and 6.9 summarise the excitonic wave function corresponding to the minimised variational energies of Fig. 6.7. The binding energy is essentially calculated over a range of discrete values of both λ and (in this case) β, hence the accuracy of these parameters is known only to the resolution of the mesh. The most important result of the variational calculation is the energy, which around the minima is a weak function of the parameters, and thus even a coarse mesh would have little consequences for the binding energy; Fig. 6.5 illustrates this nicely. The minima can be determined to $0.001\,\mathrm{meV}$ with a λ resolution of $1\,$Å and a β resolution of 0.01. In many ways the functional dependencies of λ on well width and barrier height mirror that of the binding energy E_{X^0}—it is a non-monotonic function of well width, decreases with increasing barrier height and tends towards its bulk value of $115\,$Å at large well widths.

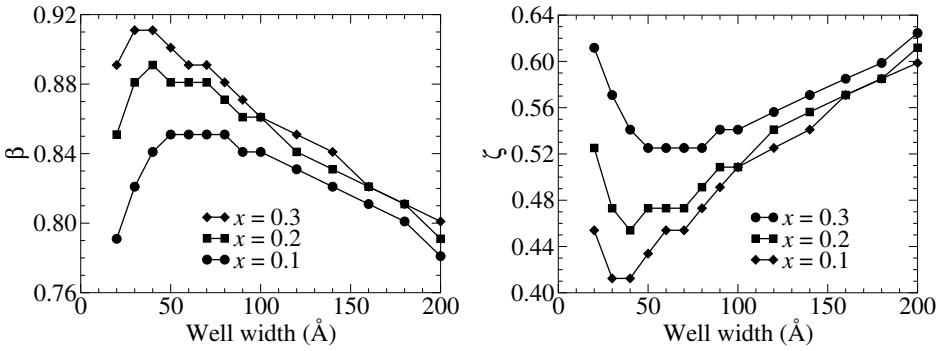


Figure 6.9: The symmetry parameters β and ζ $(= \sqrt{1-\beta^2})$ in a GaAs quantum well surrounded by $Ga_{1-x}Al_xAs$ barriers

The symmetry parameter is displayed in Fig. 6.9 as both β (as calculated) and as its transformation into ζ $(= \sqrt{1-\beta^2})$. Recalling that the trial wave function of the relative motion is given by:

$$\psi_{\mathrm{r}} = \exp\left[-\frac{\sqrt{r_\perp^2 + \zeta^2(z_{\mathrm{e}} - z_{\mathrm{h}})^2}}{\lambda}\right] \tag{6.95}$$

then $\zeta \sim$0.5 implies that:

$$\psi_{\mathrm{r}} \sim \exp\left[-\frac{\sqrt{r_\perp^2 + \frac{(z_{\mathrm{e}} - z_{\mathrm{h}})^2}{4}}}{\lambda}\right] \tag{6.96}$$

which is roughly midway between the two-dimensional case:

$$\psi_{\mathrm{r}} \sim \exp\left(-\frac{r_\perp}{\lambda}\right) \tag{6.97}$$

and the three-dimensional case:

$$\psi_{\mathrm{r}} \sim \exp\left[-\frac{\sqrt{r_\perp^2 + (z_{\mathrm{e}} - z_{\mathrm{h}})^2}}{\lambda}\right] \tag{6.98}$$

While the implications for the binding energy, using either one of the simpler wave functions above, might be relatively small, i.e. probably less than $1\,\mathrm{meV}$, the fact remains that excitons in quantum wells do have mixed symmetry. In this example the ζ value implies that 'equi-surfaces' of ψ_{r} are prolate spheroids, i.e. they look like rugby footballs (American footballs), with the major axis along the growth (z-)direction of the heterostructure.

6.7 Excitons in multiple quantum wells

Semiconductor heterostructures are often more complex than single quantum wells. Perhaps the next stage in complexity of design is to incorporate many identical wells within the same

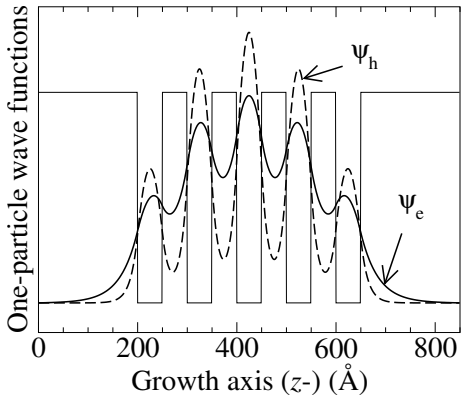


Figure 6.10: The electron and heavy-hole one-particle wave functions in a five-period 50 Å GaAs/50 Å $Ga_{0.9}Al_{0.1}As$ multiple quantum well

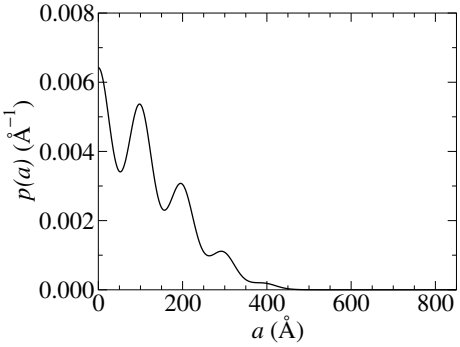


Figure 6.11: $p(a)$ for the five-period GaAs–$Ga_{0.9}Al_{0.1}As$ multiple quantum well

grown layer. As discussed previously, if the barriers separating the wells are such that the one-particle wave functions in each well overlap, then the system is called a superlattice, while if they do not overlap, it is called a multiple quantum well system.

Figure 6.10 displays the electron and heavy-hole one-particle wave functions for a five-period 50 Å GaAs/50 Å $Ga_{0.9}Al_{0.1}As$ multiple quantum well. Clearly, this system is a finite superlattice. The form of the wave functions in these finite superlattices has been discussed earlier in Chapter 3; however, a simple comparison of the two indicates that the heavy-hole wave function is more localised than that of the electron on account of its larger mass.

Although it serves no *real* purpose, it is interesting to compare the functional form of $p(a)$ with that of the infinite well of earlier. The troughs in Fig. 6.11 indicate that there are electron–hole separations which are less favoured than others. These troughs at 50, 150 and 250 Å correspond to separations where one particle would be located in a well and the other in a barrier. As in the infinite well case, the electron–hole separation (a) with the highest

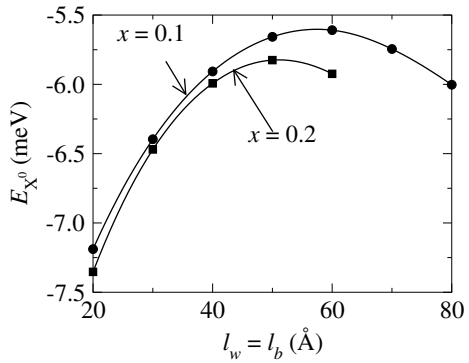


Figure 6.12: The exciton binding energy E_{X^0} as a function of well width in a five-period GaAs–$Ga_{1-x}Al_xAs$ multiple quantum well

probability is zero.

The exciton binding energy is displayed in Fig. 6.12 for a series of five-period multiple quantum wells as a function of the equal well (l_w) and barrier (l_b) widths. In contrast to the finite well, the magnitude of E_{X^0} passes through a *minimum*, which in this material system is at around 50–60 Å, before increasing again. In the GaAs/$Ga_{1-x}Al_xAs$ material system here, larger well and barrier widths reduced the interaction between the states in adjacent wells to such an extent that the separation between the symmetric ground state (as displayed in Fig. 6.10) and the first excited state was less than 10^{-6} meV, and could not be resolved; hence the truncation of the data at 60 and 80 Å. Such calculations have been performed at larger well and barrier widths in the CdTe/$Cd_{1-x}Mn_xTe$ system [8], where it was shown that beyond the minimum in the magnitude of E_{X^0} there is a maximum and then the binding energy of the multiple quantum well tends towards that of the finite well, as would be expected.

6.8 Stark ladders

As discussed previously (Chapter 3, in the context of unipolar (single charge carrier) systems), when an electric field is applied along the growth (z-)axis of a (finite or infinite) superlattice, the eigenstates which previously extended over the whole system begin to localise. As the field increases, the miniband breaks up and a localised state forms in each quantum well. At this point, the energy separation between the states in adjacent wells is proportional to the electric field and is given simply by the difference in potential energy of the well centres, i.e.

$$E_{n+1} - E_n = eFL \tag{6.99}$$

where F is the electric field strength and L is the superlattice period. The conduction- and valence-band states thus form a series of steps similar to a 'ladder'.

In a bipolar system the hole wave functions tend to localise more rapidly than those of the electron, hence a regime exists where a hole wave function centred in one particular well has a significant overlap with the electron wave functions from adjacent wells, in addition to

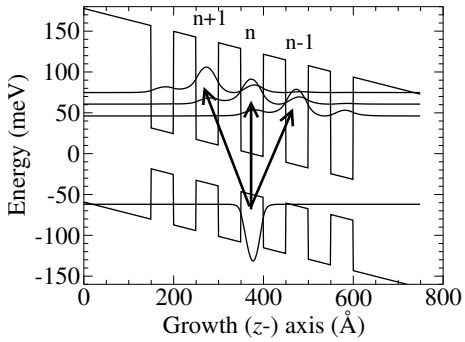


Figure 6.13: Illustration of the field ($14\,\mathrm{kV\,cm^{-1}}$) localised heavy-hole state recombining with the three closest electron states, which characterise the Stark ladder

its own. Therefore, the photoluminescence emission and absorption spectra show a series of lines, which represent exciton transitions from the hole in well n with electrons in well n, and in wells $n-1$ and $n+1$, as displayed in Fig. 6.13.

Given that the electron and hole energy separations between adjacent wells are proportional to the field, then the electron–hole separation (ignoring the exciton binding energy for now) will also be proportional to the field, as displayed in Fig. 6.14 (left). The addition of the exciton binding energy in this simplistic explanation complicates the situation slightly (see Fig. 6.14 (right)), but by the time that moderate electric fields are present, the spectral lines are equally spaced and radiating from the zero-field point. Mendez *et al*. [24] have observed very similar behaviour in a related system, namely a 30 Å GaAs/35 Å $Ga_{0.65}Al_{0.35}As$ superlattice.

To date, no excitonic states have been calculated with the variable-symmetry parameter $\zeta > 1$. However, for the work discussed in this present section, some of the Stark ladder excitons have β values tending to zero, i.e. $\zeta = 1$.

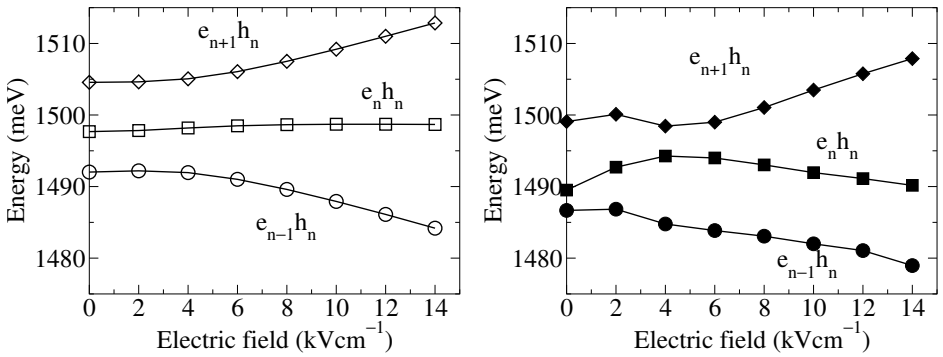


Figure 6.14: The excitonic transition energies characteristic of a Stark ladder: (left) without exciton binding energies; (right) with exciton binding energies

6.9 Self-consistent effects

In the analysis put forward in this present chapter and for the majority of the relevant literature, the exciton binding energy has been calculated by using a variational approach, based on the choice of trial wave function of the following form:

$$\Psi = \psi_{\mathrm{e}}(z_{\mathrm{e}})\psi_{\mathrm{h}}(z_{\mathrm{h}})\psi_{\mathrm{r}}(x, y, z_{\mathrm{e}} - z_{\mathrm{h}}) \tag{6.100}$$

where most importantly, ψ_{e} and ψ_{h} were simply the one-particle eigenstates of the one-dimensional Schrödinger equation of the heterostructure; this means that ψ_{e} was calculated without the presence of the hole and ψ_{h} was calculated without the presence of the electron.

Attempts to improve upon this began with the simple idea of the Coulombic hole model. This has been implemented [25] in the dilute magnetic semiconductor system CdTe/$\mathrm{Cd}_{1-x}\mathrm{Mn}_x\mathrm{Te}$ (see Section 5.11.1), which is thought to have a magnetic-field-induced Type-I to Type-II transition. It is generally accepted that such corrections to the one-particle wave functions are more important in Type-II systems, as a well-confined electron may be able to 'suck in' the hole from the adjacent layer, thus having a large effect on the exciton binding energy.

A fully self-consistent exciton model has been proposed by Warnock *et al*. [26] and later extended to include the variable-symmetry relative motion term of this analysis [27]. The self-consistent correction to the one-particle eigenfunctions, ψ_{e} and ψ_{h}, is incorporated by solving a new Schrödinger equation:

$$\left(\int_{\text{all space}} \psi_{\mathrm{h}}^{*}\psi_{\mathrm{r}}^{*}\mathcal{H}\psi_{\mathrm{h}}\psi_{\mathrm{r}}\,\mathrm{d}z_{\mathrm{h}}\,\mathrm{d}x\,\mathrm{d}y\right)\psi_{\mathrm{e}} = E\left(\int_{\text{all space}} \psi_{\mathrm{h}}^{*}\psi_{\mathrm{r}}^{*}\psi_{\mathrm{h}}\psi_{\mathrm{r}}\,\mathrm{d}z_{\mathrm{h}}\,\mathrm{d}x\,\mathrm{d}y\right)\psi_{\mathrm{e}} \tag{6.101}$$

where $\mathcal{H}$ is the total Hamiltonian describing the system, as in equation (6.7). This yields the first improved iteration to the electron wave function, $\psi_{\mathrm{e}}^{(1)}$, say. The latter is then employed to improve the hole wave function, as follows:

$$\left(\int_{\text{all space}} \psi_{\mathrm{e}}^{(1)*}\psi_{\mathrm{r}}^{*}\mathcal{H}\psi_{\mathrm{h}}\psi_{\mathrm{r}}\,\mathrm{d}z_{\mathrm{h}}\,\mathrm{d}x\,\mathrm{d}y\right)\psi_{\mathrm{h}} = E\left(\int_{\text{all space}} \psi_{\mathrm{e}}^{(1)*}\psi_{\mathrm{r}}^{*}\psi_{\mathrm{h}}\psi_{\mathrm{r}}\,\mathrm{d}z_{\mathrm{h}}\,\mathrm{d}x\,\mathrm{d}y\right)\psi_{\mathrm{h}} \tag{6.102}$$

Following this, both $\psi_{\mathrm{e}}^{(1)}$ and $\psi_{\mathrm{h}}^{(1)}$ are utilised in the standard exciton-binding-energy calculation to give the new relative motion term, characterised entirely in terms of the parameters λ and ζ. The procedure is repeated until the total energy of the exciton is minimised and λ and ζ have converged.

Piorek *et al*. [27] demonstrated that for Type-I systems, a variable-symmetry relative motion term is *more important* than self-consistency with simple 2D or 3D wave functions.

The self-consistent correction proved itself to be important in flat-band (i.e. a zero offset in the valence band) and Type-II systems. The simple one-particle solutions are usually localised in different semiconductor layers, thus leading to a quite small exciton binding energy. However, the addition of self-consistency allowed the electron and hole eigenfunctions to move closer together, thus increasing the binding energy and, most importantly, reducing the total energy of the exciton.

6.10 2s exciton

In a similar manner to both the cases of the hydrogen atom and impurities, excitons can also have excited states. While the electron and hole one-particle states remain unchanged, the corresponding choice for the relative motion factor of the 2s excited state is given by:

$$\psi_{\mathrm{r}}^{2\mathrm{s}} = \left(1 - \frac{\alpha r'}{\lambda_{2\mathrm{s}}}\right) \exp\left(-\frac{r'}{\lambda_{2\mathrm{s}}}\right) \tag{6.103}$$

where $\lambda_{2\mathrm{s}}$ is stated specifically to imply the Bohr radius of the 2s state and α is an (as yet undefined) parameter, chosen to ensure orthogonality of the 1s and 2s eigenstates, for example, as follows:

$$\mathcal{E} = \left\langle \Psi^{1\mathrm{s}} | \Psi^{2\mathrm{s}} \right\rangle = 0 \tag{6.104}$$

In the spirit of the above, this can be written as:

$$\mathcal{E} = \int_0^{\nu-\mu} p(a) L(a)\, \mathrm{d}a \tag{6.105}$$

where, in a similar manner to equation (6.18):

$$L(a) = \int \psi_{\mathrm{r}}^{1\mathrm{s}} \psi_{\mathrm{r}}^{2\mathrm{s}}\, \mathrm{d}x\, \mathrm{d}y \tag{6.106}$$

Using the specific forms for $\psi_{\mathrm{r}}^{1\mathrm{s}}$ and $\psi_{\mathrm{r}}^{2\mathrm{s}}$ as above, then:

$$L(a) = 2\pi \left\{ \sqrt{1-\beta^2} a\lambda' + \lambda'^2 - \frac{\alpha}{\lambda_{2\mathrm{s}}} \left[(1-\beta^2) a^2 \lambda' + 2\sqrt{1-\beta^2} a\lambda'^2 + 2\lambda'^3 \right] \right\} \times \exp\left(-\frac{\sqrt{1-\beta^2} a}{\lambda'}\right) \tag{6.107}$$

where $1/\lambda' = 1/\lambda_{1\mathrm{s}} + 1/\lambda_{2\mathrm{s}}$. Writing this as $L(a) = L_1 - \alpha L_2$ leads to:

$$\alpha = \frac{\int_0^{\nu-\mu} p(a) L_1(a)\, \mathrm{d}a}{\int_0^{\nu-\mu} p(a) L_2(a)\, \mathrm{d}a} \tag{6.108}$$

The remaining functions, $F(a)$, $G(a)$, $J(a)$ and $K(a)$, can all be derived by using similar procedures to the 1s case, and after lengthy manipulation are given by:

$$F(a) = 2\pi \left\{ \frac{\sqrt{1-\beta^2} a\lambda_{2\mathrm{s}}}{2} + \frac{\lambda_{2\mathrm{s}}^2}{4} - \frac{2\alpha}{\lambda_{2\mathrm{s}}} \left[\frac{(1-\beta^2) a^2 \lambda_{2\mathrm{s}}}{2} + \frac{\sqrt{1-\beta^2} a\lambda_{2\mathrm{s}}^2}{2} + \frac{\lambda_{2\mathrm{s}}^3}{4} \right] + \frac{\alpha^2}{\lambda_{2\mathrm{s}}^2} \left(\frac{(1-\beta^2)^{\frac{3}{2}} a^3 \lambda_{2\mathrm{s}}}{2} + \frac{3(1-\beta^2) a^2 \lambda_{2\mathrm{s}}^2}{4} + \frac{6\sqrt{1-\beta^2} a\lambda_{2\mathrm{s}}^3}{8} + \frac{6\lambda_{2\mathrm{s}}^4}{16} \right) \right\}$$

$$\times \exp\left(-\frac{2\sqrt{1-\beta^2}a}{\lambda_{2s}}\right) \tag{6.109}$$

In addition, $G(a) = G_1 + G_2$ where:

$$G_1 = 2\pi \int_0^1 (\alpha+1)^2 \frac{(1-\beta^2)^2 a^2}{\lambda_{2s}^2} \exp\left[-\frac{\sqrt{1-\beta^2}a}{\lambda_{2s}}\left(\frac{1}{w}+w\right)\right] \times \left[\frac{1-w^2}{w(1+w^2)}\right] \mathrm{d}w \tag{6.110}$$

and

$$G_2 = 2\pi \left[-\frac{\alpha(\alpha+1)}{\lambda_{2s}^2} + \frac{\alpha^2}{\lambda_{2s}^3}\left(\frac{\sqrt{1-\beta^2}a}{2} + \frac{\lambda_{2s}}{4}\right)\right] \times \exp\left(-\frac{2\sqrt{1-\beta^2}a}{\lambda_{2s}}\right)(1-\beta^2)^2 a^2 \tag{6.111}$$

Continuing for the remaining functions:

$$J(a) = 2\pi \int_0^1 \left[1 - \frac{\alpha\sqrt{1-\beta^2}a\left(\frac{1}{w}+w\right)}{2\lambda_{2s}}\right] \times \left\{\left[\frac{2\alpha+1}{\lambda_{2s}^2} - \frac{\alpha\sqrt{1-\beta^2}a\left(\frac{1}{w}+w\right)}{2\lambda_{2s}^3}\right]\left(\frac{1-w^2}{1+w^2}\right)^4 - \left[\frac{2(\alpha+1)w}{\lambda_{2s}\sqrt{1-\beta^2}a(1+w^2)} - \frac{\alpha}{\lambda_{2s}^2}\right]\left[1+\frac{4w^2}{(1+w^2)^2}\right]\right\} \times \exp\left[-\frac{\sqrt{1-\beta^2}a\left(\frac{1}{w}+w\right)}{\lambda_{2s}}\right]\left[\frac{(1-\beta^2)a^2}{4}\right]\left(\frac{1}{w}-w\right)\left(\frac{1}{w^2}+1\right)\mathrm{d}w \tag{6.112}$$

and finally:

$$K(a) = 2\pi\frac{\beta a}{2}\int_0^{\left(\frac{1}{\beta}-\frac{1}{\beta}\sqrt{1-\beta^2}\right)} \left[1 - \frac{\alpha\beta a}{2\lambda_{2s}}\left(\frac{1}{w}-w\right)\right]^2 \times \exp\left[-\frac{\beta a}{\lambda_{2s}}\left(\frac{1}{w}-w\right)\right]\left(\frac{1}{w^2}-1\right)\mathrm{d}w \tag{6.113}$$

Exercises

(1). Take a purely classical view of an exciton, similar to the Earth–Moon system, for example. Calculate the radius of the orbits of the electron and hole around the 'centre of gravity' for the two examples in Section 6.1, i.e. excitons in bulk GaAs and bulk CdTe.

(2). Taking the light-hole mass in GaAs and CdTe as $0.08m_0$ and $0.18m_0$ respectively, calculate the energy difference between the light- and heavy-hole excitons in both materials.

(3). With the variable symmetric trial wave function for the electron–hole interaction as in equations (6.45) and (6.46), i.e.

$$\psi_{\mathrm{r}} = \exp\left(-\frac{r'}{\lambda}\right), \qquad \text{with } r'^2 = (x_{\mathrm{e}} - x_{\mathrm{h}})^2 + (y_{\mathrm{e}} - y_{\mathrm{h}})^2 + \zeta^2(z_{\mathrm{e}} - z_{\mathrm{h}})^2 \quad (6.114)$$

a value of $\zeta = 1$ gives a spherically symmetric wave function, which would project on the y–z plane as a circle. Draw the ellipses representing the surfaces of constant value of the wave function ψ_{r} corresponding to values of $\zeta = 0.9, 0.8, 0.7, 0.6$ and 0.5.

(4). Calculate the integrals $F(a)$, $G(a)$, $J(a)$ and $K(a)$, defined in equations (6.18), (6.34), (6.40) and (6.42) respectively, for a two-dimensional wave function representing the electron–hole interaction, i.e. choose:

$$\psi_{\mathrm{r}} = \exp\left(-\frac{r_\perp}{\lambda}\right), \qquad \text{where } r_\perp^2 = (x_{\mathrm{e}} - x_{\mathrm{h}})^2 + (y_{\mathrm{e}} - y_{\mathrm{h}})^2 \quad (6.115)$$

Hence adapt the exciton-binding-energy program `qwwad_exciton_binding` to calculate the binding energy for a two-dimensional (planar) exciton.

(5). Repeat the previous question but this time for a spherically symmetric electron–hole wave function of the form:

$$\psi_{\mathrm{r}} = \exp\left(-\frac{r}{\lambda}\right), \qquad \text{where } r^2 = (x_{\mathrm{e}} - x_{\mathrm{h}})^2 + (y_{\mathrm{e}} - y_{\mathrm{h}})^2 + (z_{\mathrm{e}} - z_{\mathrm{h}})^2 \quad (6.116)$$

References

[1] R. T. Weidner and R. L. Sells, *Elementary Modern Physics*, Allyn and Bacon, Boston, Third edition, 1980.

[2] R. M. Eisberg, *Fundamentals of Modern Physics*, John Wiley & Sons, Inc., New York, 1961.

[3] S. Adachi, *GaAs and Related Materials: Bulk Semiconducting and Superlattice Properties*, World Scientific, Singapore, 1994.

[4] M. Avon and J. S. Prener, Eds., *Physics and Chemistry of II-VI Compounds*, North-Holland, Amsterdam, 1967.

[5] C. P. Hilton, W. E. Hagston, and J. E. Nicholls, 'Variational methods for calculating exciton binding energies in quantum well structures', *J. Phys. A: Math. Gen.*, **25**(8):2395, 1992.

[6] P. Hilton, J. Goodwin, P. Harrison, and W. E. Hagston, 'Theory of exciton energy levels in multiply periodic systems', *J. Phys. A: Math. Gen.*, **25**(20):5365, 1992.

[7] P. Harrison, J. P. Goodwin, and W. E. Hagston, 'Exciton energy levels and band-offset determination in magnetic superlattices', *Phys. Rev. B*, **46**:12377–12383, 1992.

[8] P. Harrison, T. Piorek, W. E. Hagston, and T. Stirner, 'The symmetry of the relative motion of excitons in semiconductor heterostructures', *Superlatt. Microstruct.*, **20**:45–57, 1996.

[9] P. Harrison, W. E. Hagston, and T. Stirner, 'Excitons in diffused quantum wells', *Phys. Rev. B*, **47**:16404–16409, 1993.

[10] R. G. Roberts, P. Harrison, T. Stirner, and W. E. Hagston, 'Stark ladders in strongly coupled finite superlattices', *J. Physique IV*, **3-C5**:203–206, 1993.

[11] P. Harrison and W. E. Hagston, 'The effect of linear and non-linear diffusion on exciton energies in quantum wells', *J. Appl. Phys.*, **79**:8451–8455, 1996.

[12] G. Bastard, E. E. Mendez, L. L. Chang, and L. Esaki, 'Exciton binding energy in quantum wells', *Phys. Rev. B*, **26**:1974, 1982.

[13] S.-K. Chang, A. V. Nurmikko, J.-W. Wu, L. A. Kolodziejski, and R. L. Gunshor, 'Band offsets and excitons in CdTe/(Cd,Mn)Te quantum wells', *Phys. Rev. B*, **37**(3):1191–1198, 1988.

[14] M. M. Dignam and J. E. Sipe, 'Exciton states in type-I and type-II GaAs/$Ga_{1-x}Al_xAs$ superlattices', *Phys. Rev. B*, **41**(5):2865–2878, 1990.

[15] U. Ekenberg and M. Altarelli, 'Exciton binding-energy in a quantum-well with inclusion of valence band coupling and nonparabolicity', *Phys. Rev. B*, **35**:7585, 1987.

[16] E. L. Ivchenko, A. V. Kavokin, V. P. Kochereshko, G. R. Posina, I. N. Uraltsev, D. R. Yakovlev, R. N. Bicknell-Tassius, A. Waag, and G. Landwehr, 'Exciton oscillator strength in magnetic-field-induced spin superlattices CdTe/(Cd,Mn)Te', *Phys. Rev. B*, **46**(12):7713–7722, 1992.

[17] A. Bellabchara, P. Lefebvre, P. Christol, and H. Mathieu, 'Improved modeling of excitons in type-II semiconductor heterostructures by use of a three-dimensional variational function', *Phys. Rev. B*, **50**:11840, 1994.

[18] J. Cen, S. V. Branis, and K. K. Bajaj, 'Exciton binding energies in finite-barrier type-II quantum-well structures in a magnetic field', *Phys. Rev. B*, **44**(23):12848–12856, 1991.

[19] Y. Shinozuka and M. Matsuura, 'Wannier excitons in quantum wells', *Phys. Rev. B*, **28**:4878, 1983.

[20] S. V. Branis, J. Cen, and K. K. Bajaj, 'Effect of magnetic fields on exciton binding energies in type-II GaAs–AlAs quantum-well structures', *Phys. Rev. B*, **44**(20):11196–11202, 1991.

[21] Y.-P. Feng, H. S. Tan, and H. N. Spector, 'Quantum well excitons in an electric field: two versus three dimensional behavior', *Superlatt. Microstruct.*, **17**(3):267, 1995.

[22] J. W. Wu and A. V. Nurmikko, 'Wannier excitons in semiconductor quantum wells with small valence band-offsets—a generalised variational approach', *Phys. Rev. B*, **38**:1504, 1988.

[23] M. J. L. S. Haines, N. Ahmed, S. J. A. Adams, K. Mitchell, I. R. Agool, C. R. Pidgeon, B. C. Cavenett, E. P. O'Reilly, A. Ghiti, and M. T. Emeny, 'Exciton-binding-energy maximum in $Ga_{1-x}In_xAs$/GaAs quantum wells', *Phys. Rev. B*, **43**(14):11944–11949, 1991.

[24] E. E. Mendez, F. Agulló-Rueda, and J. M. Hong, 'Stark localization in GaAs–GaAlAs superlattices under an electric field', *Phys. Rev. Lett.*, **60**(23):2426–2429, 1988.

[25] A. V. Kavokin, V. P. Kochereshko, G. R. Posina, I. N. Uraltsev, D. R. Yakovlev, G. Landwehr, R. N. Bicknell-Tassius, and A. Waag, 'Effect of the electron Coulomb potential on hole confinement in II–VI quantum wells', *Phys. Rev. B*, **46**(15):9788–9791, 1992.

[26] J. Warnock, B. T. Jonker, A. Petrou, W. C. Chou, and X. Liu, 'Exciton energies in shallow quantum wells and spin superlattices', *Phys. Rev. B*, **48**(23):17321–17330, 1993.

[27] T. Piorek, W. E. Hagston, and P. Harrison, 'Relative importance of self-consistency and variable symmetry in the calculation of exciton energies in type-I and type-II semiconductor heterostructures', *Phys. Rev. B*, **52**(19):14111–14117, 1995.

7

Strained quantum wells

V. D. Jovanović

University of Leeds

7.1 Stress and strain in bulk crystals

A mechanical force acting on a crystal lattice changes the relative positions of the lattice points (sites) i.e. the positions of the atoms forming the crystal structure. This can be characterised by a vector $\mathbf{u}$, which defines the relative displacement of an atom into a new position $\mathbf{r}' = \mathbf{r} + \mathbf{u}$ in some arbitrary Cartesian coordinate system. Different crystal lattice points can have different relative displacements making the vector $\mathbf{u}$ coordinate dependent, i.e. $\mathbf{u} = \mathbf{u}(\mathbf{r})$. If the lattice points return to their original positions after the force is removed, then the deformation is described as elastic.

Knowing the relative displacements of each lattice site, the state of the crystal deformation can be described by strain components defining a second-rank tensor (for more about application of tensors in crystals, see Nye [1]) as:

$$\epsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \qquad i,j = 1,2,3 \tag{7.1}$$

where u_1, u_2, and u_3 are the relative displacements of the crystal lattice points along the x_1-, x_2- and x_3-axes, respectively. The diagonal components represent extensions per unit length along the x_1-, x_2- and x_3-directions and are usually referred as 'stretches' (see Fig. 7.1), while the off-diagonal components ϵ_{ij} are related to 'rotations', e.g. the term $\partial u_i/\partial x_j$ represents a rotation about the x_3-axis towards x_1 of a line element parallel to x_2. As such, the angle between the two line elements parallel to x_1 and x_2 changes from $\pi/2$ before deformation to $\pi/2 - \epsilon_{ij}$ after.[1] The strain tensor is symmetrical ($\epsilon_{ij} = \epsilon_{ji}$) and can be

[1]Note that in the literature an intuitive notation is occasionally seen replacing the indices 1, 2 and 3 with x, y and z, e.g. $\epsilon_{11} \rightarrow \epsilon_{xx}$.

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

written as:

$$\epsilon = \begin{pmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{31} \\ \epsilon_{12} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{23} & \epsilon_{33} \end{pmatrix} \tag{7.2}$$

If the shear strain components are zero, the diagonal elements also determine the change in the crystal volume (also known as 'dilation') as:

$$\frac{\Delta V}{V} = \text{Tr}(\epsilon) = \epsilon_{11} + \epsilon_{22} + \epsilon_{33} \tag{7.3}$$

This can be explained in terms of the unity cube strained by the diagonal components. Assuming the strains are small, which is true in the limits of linear strain theory as well as in most applications, the distorted volume will then be $(1 + \epsilon_{11})(1 + \epsilon_{22})(1 + \epsilon_{33})$, giving a total change in the cube volume (to first order) of $\epsilon_{11} + \epsilon_{22} + \epsilon_{33}$ (see Fig. 7.1).

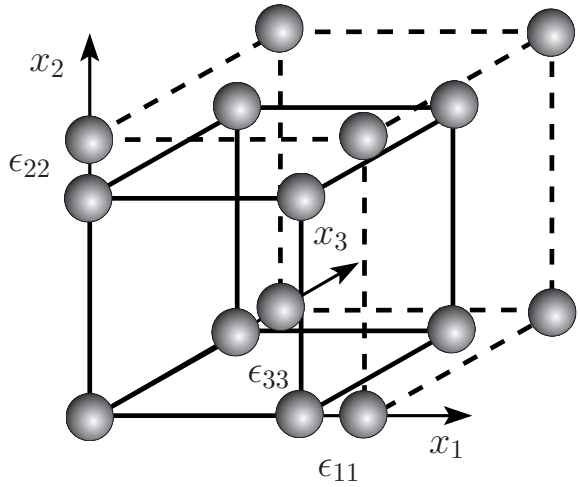


Figure 7.1: Schematic illustration of the influence of the diagonal strain components or stretches (ϵ_{11}, ϵ_{22}, ϵ_{33}) on the crystal lattice

It is also important to define the properties of the force causing the strain in the crystal lattice. If a crystal is acted on by an external force or if part of a crystal is applying a force on a neighbouring part, then the crystal is said to be in the state of 'stress', where the stress is usually defined as the force per unit area of the crystal. If a unit cube is considered, then normal stress components can be defined as σ_{ii}, i.e. σ_{11}, σ_{22}, σ_{33}, and the shear stress components as σ_{ij}, e.g. σ_{12}, σ_{21}, σ_{13}, etc., as shown in Fig. 7.2. In the homogeneous case when the force is independent of the position on the crystal surface, the stress components form a symmetric second-rank tensor ($\sigma_{ij} = \sigma_{ji}$) as:

$$\sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{31} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{23} & \sigma_{33} \end{pmatrix} \tag{7.4}$$

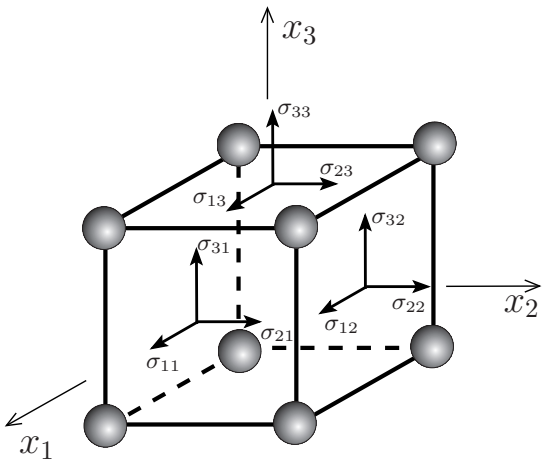


Figure 7.2: Directions of the stress components acting on a crystal lattice that are usually assumed

The diagonal elements of the stress tensor are of special importance in applications. If only σ_{11} is a non-zero stress component, then the stress is said to be 'uniaxial', while if both σ_{11} and σ_{22} are not equal to zero the stress is defined as 'biaxial'. The latter is the case common in quantum well structures as elaborated on in the following section.

In the limit of linear elastic theory the stress and the strain are connected by Hooke's law, which reads:

$$\epsilon_{ij} = \sum_{k=1}^{3}\sum_{l=1}^{3} S_{ijkl}\sigma_{kl} \tag{7.5}$$

where S_{ijkl} are the 'elastic compliance constants' which form a fourth-rank tensor. In the literature, stress is commonly presented as a function of strain and Hooke's law takes the similar form:

$$\sigma_{ij} = \sum_{k=1}^{3}\sum_{l=1}^{3} C_{ijkl}\epsilon_{kl} \tag{7.6}$$

where C_{ijkl} are the 'elastic stiffness constants' and also form a fourth-rank tensor.

A more concise matrix (or Voigt's) notation is commonly used in the literature which reduces the number of tensor indices and transforms the strain, stress and elastic stiffness components as follows:

$$\begin{pmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{31} \\ \epsilon_{12} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{23} & \epsilon_{33} \end{pmatrix} \rightarrow \begin{pmatrix} \epsilon_1 & \frac{1}{2}\epsilon_6 & \frac{1}{2}\epsilon_5 \\ \frac{1}{2}\epsilon_6 & \epsilon_2 & \frac{1}{2}\epsilon_4 \\ \frac{1}{2}\epsilon_5 & \frac{1}{2}\epsilon_4 & \epsilon_3 \end{pmatrix} \tag{7.7}$$

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{31} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{23} & \sigma_{33} \end{pmatrix} \rightarrow \begin{pmatrix} \sigma_1 & \sigma_6 & \sigma_5 \\ \sigma_6 & \sigma_2 & \sigma_4 \\ \sigma_5 & \sigma_4 & \sigma_3 \end{pmatrix} \tag{7.8}$$

$$C_{ijkl} \rightarrow C_{mn}, \qquad i,j,k,l = 1,2,3;\ m,n = 1,\ldots,6 \tag{7.9}$$

The $\frac{1}{2}$ terms have been introduced in order to give a clear relationship between the stress, strain and other parameters used. Hooke's law given by equation (7.6) in the matrix notation takes a simplified form and reads:

$$\sigma_i = \sum_{k=1}^{6} C_{ik}\epsilon_k \tag{7.10}$$

or equivalently:

$$\begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \\ \sigma_5 \\ \sigma_6 \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{14} & C_{15} & C_{16} \\ C_{21} & C_{22} & C_{23} & C_{24} & C_{25} & C_{26} \\ C_{31} & C_{32} & C_{33} & C_{34} & C_{35} & C_{36} \\ C_{41} & C_{42} & C_{43} & C_{44} & C_{45} & C_{46} \\ C_{51} & C_{52} & C_{53} & C_{54} & C_{55} & C_{56} \\ C_{61} & C_{62} & C_{63} & C_{64} & C_{65} & C_{66} \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix} \tag{7.11}$$

where C_{ik} is now the elastic stiffness matrix, which is related to the elastic compliance matrix as $[C] = [S]^{-1}$ and defines the unique correlation between the elastic constants.

In order to define the elastic properties of the crystal, 36 independent elastic stiffness constants are required. However, crystal lattices commonly exhibit certain symmetries, which can be employed to reduce the number of constants necessary to describe their elastic behaviour. The numbers of independent matrix components for the most used semiconductors are: 3 for the cubic (e.g. GaAs) and 5 for the hexagonal (e.g. wurtzite GaN) crystal geometries. Hence, the C matrices for these semiconductors are given as:

$$\begin{bmatrix} C_{11} & C_{12} & C_{12} & 0 & 0 & 0 \\ C_{12} & C_{11} & C_{12} & 0 & 0 & 0 \\ C_{12} & C_{12} & C_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & C_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & C_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & C_{44} \end{bmatrix} \tag{7.12}$$

and

$$\begin{bmatrix} C_{11} & C_{12} & C_{13} & 0 & 0 & 0 \\ C_{12} & C_{11} & C_{13} & 0 & 0 & 0 \\ C_{13} & C_{13} & C_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & C_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & C_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2}(C_{11} - C_{12}) \end{bmatrix} \tag{7.13}$$

for the cubic and the hexagonal crystals, respectively.

Finally, to deform the crystal lattice, a certain amount of energy is needed. The sum of the work done by the stress components acting on the crystal defines the strain energy as another important parameter influencing the crystal strain state. In terms of the strain components with respect to the validity of Hooke's law, the strain energy per unit volume can be expressed

in matrix notation as:

$$W = \frac{1}{2}\sum_{i=1}^{6}\sum_{j=1}^{6} C_{ij}\epsilon_i\epsilon_j \tag{7.14}$$

The strain energy density always has to be greater than zero. Furthermore, the crystal system reaches the most stable strain state for the minimal value of the strain energy—a condition which is used in the derivation of strain balancing in Section 7.4.

For cubic semiconductors, equations (7.12) and (7.14) give the elastic energy as:

$$W_{\text{cubic}} = \frac{1}{2}C_{11}\left(\epsilon_1^2+\epsilon_2^2+\epsilon_3^2\right) + \frac{1}{2}C_{44}\left(\epsilon_4^2+\epsilon_5^2+\epsilon_6^2\right) + C_{12}\left(\epsilon_1\epsilon_2+\epsilon_2\epsilon_3+\epsilon_3\epsilon_1\right) \tag{7.15}$$

whilst for hexagonal crystals it takes the form:

$$\begin{aligned} W_{\text{hex}} = {} & \frac{1}{2}C_{11}\left(\epsilon_1^2+\epsilon_2^2\right) + \frac{1}{2}C_{33}\epsilon_3^2 + C_{12}\epsilon_1\epsilon_2 + C_{13}\epsilon_3\left(\epsilon_1+\epsilon_2\right) \\ & + \frac{1}{2}C_{44}\left(\epsilon_4^2+\epsilon_5^2\right) + \frac{1}{4}(C_{11}-C_{12})\epsilon_6^2 \end{aligned} \tag{7.16}$$

7.2 Strain in quantum wells

The effects of strain are of particular interest in quantum well structures. If a thin epitaxial layer is deposited on a much thicker substrate (usually assumed to be infinitely thick in comparison to the epitaxial layer), then the lattice constant (a_l) in the growth plane (perpendicular to the growth direction) of the layer will be forced to change so as to be equal to the lattice constant of the substrate (a_0). As a consequence, the crystal lattice is under biaxial stress along the growth interface and, while no force is applied along the growth direction, the crystal is able to relax freely along that direction. Therefore, the stress has only two diagonal components σ_1 and σ_2, while σ_3 and the shear components are zero. Hence, in matrix notation the stress in the epitaxial (quantum well) layer can be written as:

$$\sigma = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{7.17}$$

The existence of biaxial stress results in the appearance of an in-plane strain. Growth that allows the lattice constant of the epitaxial layer to fully equal (match) the substrate is usually referred to as *pseudomorphic growth*. In such cases, the in-plane strain can be easily calculated as:

$$\epsilon_{\parallel} = \frac{a_0 - a_l}{a_l} \tag{7.18}$$

In quantum well systems the in-plane strain is usually of the order of 1%, e.g. for an AlN layer grown on a GaN substrate $\epsilon_{\parallel} \approx 2\%$. However, some materials and their alloys have very similar lattice constants, such as the GaAs/AlGaAs system for which $\epsilon_{\parallel} < 0.1\%$, allowing them to be considered as unstrained.

The lattice constant of the epitaxial layer can be either larger or smaller than the lattice constant of the substrate (see Fig. 7.3), defining a compressive ($\epsilon_{\parallel} < 0$) or a tensile ($\epsilon_{\parallel} > 0$)

strain, respectively. A compressive strain will therefore force the lattice constant in the plane to shrink (see material B in Fig. 7.3); conversely tensile strain will force it to expand (see material A in Fig. 7.3). It is also reasonable to assume that the strain in the thin semiconductor layers is homogeneous and therefore constant throughout the layer.

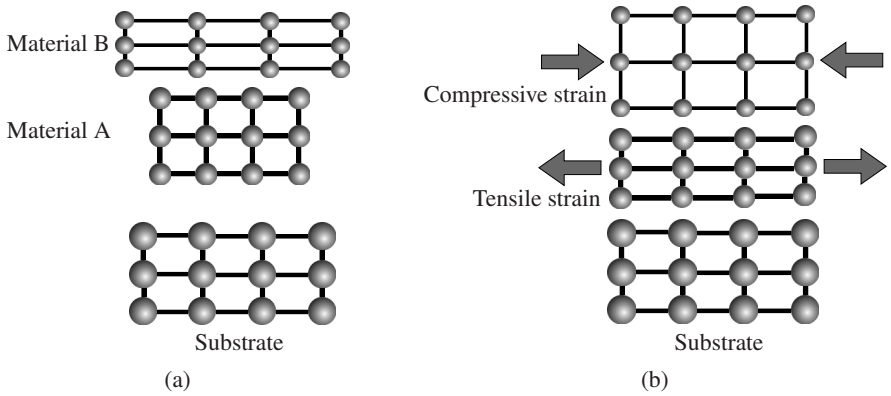


Figure 7.3: Schematic illustration of a substrate with two mismatched layers: (a) shows the free-standing unstrained layers with lattice constants smaller $a_l < a_0$ (material A) and larger $a_l > a_0$ (material B) than the lattice constant of the substrate; and (b) shows the situation when either of the materials is grown on the substrate, i.e. material A is under tensile strain and material B is under compressive strain as their lattice constants are forced to be equal to the lattice constant of the substrate

Though no stress exists ($\sigma_3 = 0$) in the growth direction, the lattice constant is still forced to change due to the 'Poisson effect' (see Fig. 7.3). If the compressive strain forces the in-plane lattice constant to reduce, then the lattice constant in the growth direction will increase, and vice versa for tensile strain. Hence, strain exists in the growth direction as well. The ratio that determines the increase or decrease of the lattice constant due to the in-plane stress is called Poisson's ratio (ν) and it connects the in-plane and the perpendicular strains as:

$$\epsilon_3 = -\nu\epsilon_1 \tag{7.19}$$

For the commonly used cubic semiconductor materials grown along the $[001]$ direction, the relationship between the stress and the strain in the epitaxial (quantum well) layers is given by Hooke's law (equations (7.6) and (7.12)). For biaxial stress ($\sigma_3 = 0$), this reads:

$$\begin{bmatrix} \sigma_1 \\ \sigma_2 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & C_{12} & 0 & 0 & 0 \\ C_{12} & C_{11} & C_{12} & 0 & 0 & 0 \\ C_{12} & C_{12} & C_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & C_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & C_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & C_{44} \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix} \tag{7.20}$$

The matrix equation defines a system of linear independent equations. The in-plane strain components (ϵ_1 and ϵ_2) are known values, defined by the lattice deformation, as in equation (7.18). The perpendicular strain ϵ_3 can be obtained from the third row of the matrix equation:

$$\begin{aligned} \epsilon_1 = \epsilon_2 &= \epsilon_{\parallel} \\ \epsilon_3 &= -2\frac{C_{12}}{C_{11}}\epsilon_1 \end{aligned} \tag{7.21}$$

The shear strain components ϵ_4, ϵ_5 and ϵ_6 are all found to be zero, as expected, from the final rows of the matrix equation. The complete strain tensor may be written in matrix form:

$$\epsilon = \begin{pmatrix} \epsilon_1 & 0 & 0 \\ 0 & \epsilon_1 & 0 \\ 0 & 0 & -2\frac{C_{12}}{C_{11}}\epsilon_1 \end{pmatrix} \tag{7.22}$$

where the factor $\nu = -2\frac{C_{12}}{C_{11}}$ represents Poisson's ratio for cubic semiconductors in the $[001]$ direction. Typical values of the strain components and the strain distribution in cubic InGaAs/AlGaAs quantum wells are presented in Fig. 7.4.

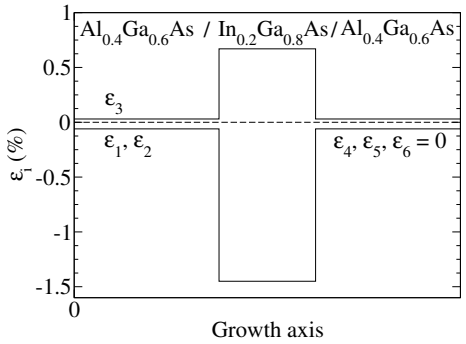


Figure 7.4: Values of the strain components in an $In_{0.2}Ga_{0.8}As$ quantum well surrounded by $Al_{0.4}Ga_{0.6}As$ barriers

Knowing the strain components makes it possible to determine the relative change in the lattice volume (or dilation) using equation (7.3) which now reads:

$$\Theta = \frac{\Delta V}{V} = 2\epsilon_1 \left(1 - \frac{C_{12}}{C_{11}}\right) \tag{7.23}$$

For quantum wells with the layers based on the hexagonal crystal geometry grown in the $[0001]$ direction a similar approach can be applied, modifying Hooke's law as:

$$\begin{bmatrix} \sigma_1 \\ \sigma_2 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & C_{13} & 0 & 0 & 0 \\ C_{12} & C_{11} & C_{13} & 0 & 0 & 0 \\ C_{13} & C_{13} & C_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & C_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & C_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2}(C_{11} - C_{12}) \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix} \tag{7.24}$$

which implies the solutions for the strain components are:

$$\epsilon_1 = \epsilon_2 = \epsilon_{\|} \quad \text{and} \quad \epsilon_3 = -2\frac{C_{13}}{C_{33}}\epsilon_1 \tag{7.25}$$

or in matrix form:

$$\epsilon = \begin{pmatrix} \epsilon_1 & 0 & 0 \\ 0 & \epsilon_1 & 0 \\ 0 & 0 & -2\frac{C_{13}}{C_{33}}\epsilon_1 \end{pmatrix} \tag{7.26}$$

Similarly as for the cubic crystal geometry, for epitaxial layers based on hexagonal materials the relative change of the volume can be expressed as:

$$\Theta = \frac{\Delta V}{V} = 2\epsilon_1 \left(1 - \frac{C_{13}}{C_{33}}\right) \tag{7.27}$$

7.3 Critical thickness of layers

Thus far,[2] it has been assumed that each epitaxial layer is sufficiently thin to deform elastically to match the lattice constant of the material. However, if the layer is thicker than some critical thickness, it is unable to deform uniformly, and instead relaxes in the plane (perpendicular to the growth direction), restoring its original lattice constant. This is a violent process producing a large number of defects and imperfections (cracks) in the growth surface. The formation of defects and the existence of a critical thickness can be understood in terms of the elastic energy. The strained layer system possesses an additional elastic energy which is a function of the layer thickness (approximately a product of the elastic energy density, the area and the layer width). If, for a specific layer width, the strain energy exceeds the energy required for the generation of defects, then the system will tend to relax to a new state with lower strain energy by forming imperfections in the growth plane. Therefore, pseudomorphic growth of high-quality heterostructures (with a small number of intrinsic defects) is typically restricted to thin layers of material.

The critical thickness h_c above which this relaxation process occurs may be obtained through an analysis of the elastic energy (see [2]):

$$h_c = \frac{b}{2\pi f} \frac{1 - \nu\cos^2\alpha}{(1+\nu)\cos\lambda} \left(1 + \ln\frac{h_c}{b}\right) \tag{7.28}$$

[2]This section was added in the fourth edition of this book by the principal authors.

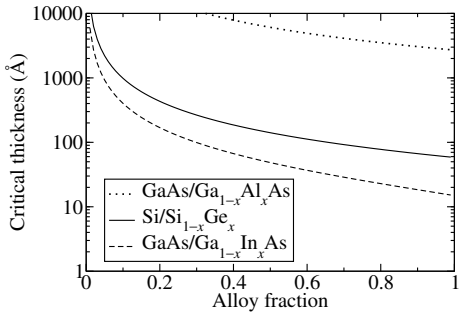


Figure 7.5: Critical thickness of films of alloy materials with variable composition upon a fixed-composition substrate

The dislocations considered in this expression are characterised by the magnitude b of their *Burgers vector* (i.e. the size of the distortion they induce in the crystal lattice), and the angles α and λ that describe the direction of the dislocation line. A critical parameter here is the misfit f between the substrate and the epitaxial layer, which is given by:

$$f = \frac{a - a_0}{a_0} \tag{7.29}$$

For clarity, the prefactor in equation (7.28) may be replaced by a constant, A, and the term in parentheses may be written as a single logarithm. This leaves an implicit equation of the form:

$$h_\mathrm{c} = A \ln\left(\frac{\exp(1)}{b} h_\mathrm{c}\right) \tag{7.30}$$

in which the unknown value of h_c appears on both sides. This can be solved iteratively using techniques such as Newton–Raphson iteration, although a direct solution also exists in terms of the special Lambert W_{-1} function [3]:

$$h_\mathrm{c} = -AW_{-1}\left(-\frac{b}{A\exp(1)}\right) \tag{7.31}$$

This special function is defined as the solution of the equation $x = W\mathrm{e}^W$ for real numbers in the range $W \leq -1$, and its value is tabulated in free software libraries such as the *GNU Scientific Library* [4].

By way of an example, many cubic crystals exhibit dislocations with a propagation direction $\alpha = \lambda = 60^\circ$ and a Burgers vector given by $\frac{a}{2}\langle 110\rangle$. The magnitude is, therefore, $b = \frac{a\sqrt{2}}{2}$, and both cosine terms in equation (7.28) evaluate to $\frac{1}{2}$. Recalling that the Poisson ratio for a biaxially strained cubic crystal is $\nu = \frac{2C_{12}}{C_{11}}$, the parameter in equation (7.30) becomes:

$$A = \frac{a}{2\sqrt{2}\pi f} \frac{1 - \frac{C_{12}}{2C_{11}}}{\frac{1}{2}\left(1 + \frac{2C_{12}}{C_{11}}\right)} \tag{7.32}$$

which simplifies to:

$$A = \frac{a}{\sqrt{2}\pi f} \frac{2C_{11} - C_{12}}{2C_{11} - 4C_{12}} \tag{7.33}$$

Figure 7.5 shows the results of this calculation for a range of alloy materials upon a substrate with fixed composition. For all materials, it is shown that $h_c \to \infty$ as $x \to 0$. This is an intuitively correct result, since the 'alloy' is the same as the substrate material in this case, and there is consequently no misfit between the lattice constants of the substrate and the layer. The layers can therefore be grown to infinite thickness. Conversely, as the alloy fraction in the layer changes, the misfit between the alloy constants increases, and the critical thickness decreases. For $Ga_{1-x}Al_xAs$ films on GaAs, the lattice constants are very similar (5.65 and 5.66 Å for GaAs and AlAs, respectively), and the system is described as being *lattice matched*. As such, very thick layers (in excess of 3000 Å) of $Ga_{1-x}Al_xAs$ can be grown upon GaAs substrates. In the case of other material systems, the lattice constants can be highly mismatched (e.g., 6.06 and 5.65 Å for InAs and GaAs, respectively), resulting in significantly reduced critical thicknesses for growth, often setting a limit of less than 100 Å for layers. Clearly, this imposes a serious practical limitation on the range of quantum heterostructure devices that can be grown in non-lattice-matched material systems.

It is important to note, however, that these calculations tend to underestimate the maximum layer thickness, since they do not account for the activation energy barrier for the formation of dislocations in a strained structure [5]. This can enable *metastable* highly strained layers to be grown to thicknesses well in excess of the critical value. Furthermore, the maximum thickness of layers can be increased by careful design of the heterostructure, or an appropriate choice of substrate material as detailed in the following section.

7.4 Strain balancing

In order to ensure pseudomorphic growth of multiple layers of quantum wells, a technique called 'strain balancing' is commonly employed. This implies a careful choice of the epitaxial layer compositions and widths as well as in the choice of the substrate (or buffer layer) parameters if possible, with the aim of engineering the strain state of the entire structure, and therefore preventing the appearance of misfit dislocations and other defects. However, the available margins for such parameter manipulation are quite small as the requirements of the quantum well structure enforce tight constraints.

The general method for evaluating a 'strain-balancing condition' is based on the fact that the stable strain state corresponds to the minimum of the elastic energy in each layer of the quantum well structure. However, when the multilayer system is considered, such a state needs to be reached across the entire structure and not only in a single layer so that a favourable strain distribution is reached overall and prevents the relaxation process. Such a condition can be met by minimising the average elastic energy with respect to the in-plane strain (known as the 'zero-stress' condition) [6] given as:

$$\bar{W} = \frac{\sum_{k=1}^{n} W_k l_k}{\sum_{k=1}^{n} l_k} \tag{7.34}$$

where l_k is the width and W_k the strain energy density of the kth layer.

Recalling the elastic energy density definition stated in equation (7.14), the energy density of the kth layer can be written as

$$W_k = \frac{1}{2}\sum_{i=1}^{6}\sum_{j=1}^{6} C_{ij}^{(k)}\epsilon_i^{(k)}\epsilon_j^{(k)} \tag{7.35}$$

where $C_{ij}^{(k)}$ is the elastic stiffness constant and $\epsilon_i^{(k)}$ is the strain component (in matrix notation) of the kth layer.

For the commonly used cubic and hexagonal semiconductors, the elastic energy densities are given by equations (7.15) and (7.16). If the nature of the strain in the epitaxial layers is assumed to be due to a biaxial stress, then the equations can be written in a simpler form by substituting equation (7.12) or (7.13) and equation (7.21) into equation (7.14) for the cubic and equivalently for the hexagonal crystals. The elastic energy density then reads:

$$W_k = A_k \cdot [\epsilon_1^{(k)}]^2 \tag{7.36}$$

where:

$$A_k = C_{11} + C_{12} - 2\frac{C_{12}^2}{C_{11}} \tag{7.37}$$

for the cubic and:

$$A_k = C_{11} + C_{12} - 2\frac{C_{13}^2}{C_{33}} \tag{7.38}$$

for the hexagonal crystal geometries.

With respect to equation (7.34), equation (7.35) can then written as:

$$\bar{W} = \frac{\sum_{k=1}^{n} A_k[\epsilon_1^{(k)}]^2 l_k}{\sum_{k=1}^{n} l_k} \tag{7.39}$$

The pseudomorphic condition implies that the modified lattice constants (a_k and a_{k-1}) of adjacent layers must be equal to that of the substrate. The in-plane strains in each of the layers can then be written as:

$$\epsilon_1^{(k)} = \frac{a_0 - a_k}{a_k} \tag{7.40}$$

and

$$\epsilon_1^{(k-1)} = \frac{a_0 - a_{k-1}}{a_{k-1}} \tag{7.41}$$

which means that the in-plane strains in adjacent layers ($\epsilon_1^{(k)}$ and $\epsilon_1^{(k-1)}$) can be related as:

$$\epsilon_1^{(k)} = \frac{a_{k-1}}{a_k}\epsilon_1^{(k-1)} + \frac{a_{k-1} - a_k}{a_k} \tag{7.42}$$

and the average in-plane stress can follow as:

$$\bar{\sigma} = \frac{\partial \bar{W}}{\partial \epsilon_1^{(1)}} = \frac{2}{l_1 + l_2 + l_3 + \ldots} \times \left\{ A_1\epsilon_1^{(1)} l_1 + A_2\epsilon_1^{(2)} l_2 \frac{\partial \epsilon_1^{(2)}}{\partial \epsilon_1^{(1)}} + A_3\epsilon_1^{(3)} l_3 \frac{\partial \epsilon_1^{(3)}}{\partial \epsilon_1^{(1)}} + \ldots \right\} \tag{7.43}$$

The zero-stress (or strain-balance) condition implies that the *in-plane* stress is zero ($\bar{\sigma} = 0$), and, using $\partial\epsilon_1^{(k)}/\partial\epsilon_1^{(1)} = a_1/a_k$ ($k = 2, 3, \ldots$), equation (7.43) gives:

$$A_1\epsilon_1^{(1)}l_1 + A_2\epsilon_1^{(2)}l_2\frac{a_1}{a_2} + A_3\epsilon_1^{(3)}l_3\frac{a_1}{a_3} + \ldots = 0 \tag{7.44}$$

which delivers the lattice constant of the substrate (or suitably grown buffer layer) necessary for strain-balancing the quantum well stack as:

$$a_0 = \frac{\sum_{k=1}^{n} A_k l_k/a_k}{\sum_{k=1}^{n} A_k l_k/a_k^2} \tag{7.45}$$

If this lattice constant is not equal to that of any readily available substrate, then it can be achieved by growth of a suitable buffer layer. For example, in the $Si_{1-x}Ge_x$/Si material system, strain balancing is achieved by growing a buffer layer with the appropriate Ge composition, i.e. $Si_{1-y}Ge_y$. This is usually linearly graded from the pure silicon composition of the substrate to the required composition y.

If a multilayer structure has periodicity, the above expression should be understood in terms of the single period, i.e. it should be applied to the n layers constituting a single period, and the whole structure will then clearly be strain-balanced.

The previous derivation takes into account the difference in elastic properties of the layers (i.e. elastic stiffness constants). If the elastic constants are similar for all layers of the structure, then the strain-balancing condition can be simplified by taking into account only the difference of the lattice constants (also known as the 'average lattice method') as:

$$a_0 = \frac{\sum_{k=1}^{n} l_k a_k}{\sum_{k=1}^{n} l_k} \tag{7.46}$$

If the substrate is fixed and cannot be engineered, the quantum well structure itself has to be 'tailored' to match the substrate, the complexity of which depends on the number of different layers per period and the boundaries imposed by the desired application. In something like a quantum well infrared photodetector this could be achieved by altering the width of the thick barrier layer separating the quantum well absorbing regions, as provided the thickness of this layer is above some minimum (to limit the dark current) its precise value is not too important. In a quantum cascade laser the injector region offers some flexibility to allow a design to be created which also satisfies the strain-balancing condition.

7.5 Effect on the band profile of quantum wells

The strain in epitaxial layers acts to change the crystal lattice geometry, i.e. to perturb its size and symmetry. This results in a change of the electronic structure; in particular a modification of the conduction- and valence-band edges and therefore a corresponding shift in the energy levels. In order to give a quantitative measure of this further effect of strain, a general case of the strained crystal is considered under the framework of deformation potential theory. This section will consider the simple case of the conduction-band Γ minimum, while more detailed analysis exploring the valence band will be presented in Chapter 14.

If the strain is assumed to be small, which is the case in semiconductor quantum wells ($\sim$1%), first-order perturbation theory can be used to calculate the band shift. The

Hamiltonian under strain can be expressed as the sum of the unperturbed Hamiltonian ($\mathcal{H}_0$) and the strain-induced contribution ($\mathcal{H}_\epsilon$), i.e.

$$\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_\epsilon \tag{7.47}$$

where the dependence of $\mathcal{H}_\epsilon$ on the strain is:

$$\mathcal{H}_\epsilon = \sum_{i,j} \frac{\partial V}{\partial \epsilon_{ij}} \epsilon_{ij} \tag{7.48}$$

and V is the original (unperturbed) crystal potential. From perturbation theory it follows that the energy shift due to strain can be expressed in terms of the deformation potential components representing the matrix elements of the relative band shift due to strain (the $\frac{\partial V}{\partial \epsilon_{ij}}$ term in equation (7.48)) as:

$$\delta E_\mathrm{c} = \sum_{i,j} D_{ij} \epsilon_{ij} \tag{7.49}$$

Note that the D_{ij} form a second-rank tensor, though the matrix representation is commonly used. The number of non-zero D_{ij} components is dependent on the crystal symmetry (similarly to strain theory presented earlier) and the type of band minimum.

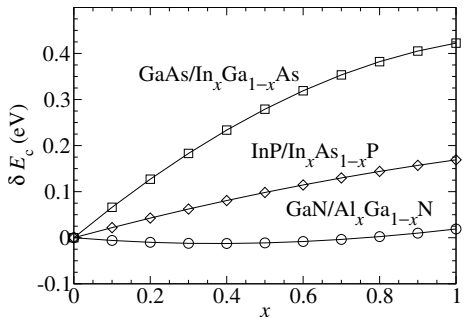


Figure 7.6: Conduction-band energy shifts for strained layers based on different materials—AlGaN grown on GaN substrate (circles), InAsP grown on InP (diamonds) and InGaAs grown on GaAs (squares)—as a function of the layer composition

For the conduction-band Γ point in cubic semiconductors all the off-diagonal deformation potential components vanish ($D_{ij} = 0$), leaving only three equal diagonal constants $D_{11} = D_{22} = D_{33} = a_\mathrm{c}$ (the isotropic case) which are usually determined by fitting to experimental data. Equation (7.49), giving the shift of the conduction-band energy due to strain, now reads:

$$\delta E_\mathrm{c}^\mathrm{cubic} = a_\mathrm{c}(\epsilon_1 + \epsilon_2 + \epsilon_3) \tag{7.50}$$

It is clear from the equation that the energy band shift depends only on the strain state of the crystal given via the diagonal strain components.

Hexagonal crystals exhibit much less symmetry than cubic ones. Consequently, a certain anisotropic behaviour can be expected and this is reflected in the different deformation potential constants. For example, GaN in the growth plane has identical deformation potential constants $D_{11} = D_{22} = a_{c\perp}$, whilst in the growth direction the constant $D_{33} = a_{c\parallel}$ is almost three times larger. The total conduction-band shift in GaN then follows as:

$$\delta E_c^{\text{hex}} = a_{c\perp}(\epsilon_1 + \epsilon_2) + a_{c\parallel}\epsilon_3 \tag{7.51}$$

The previous analysis can be applied to determine the energy shift of the conduction-band edge in the vicinity of the Γ minimum in the epitaxial layer of the quantum well structure. For the layers based on cubic semiconductor material grown in the $[001]$ direction, substituting the strain components given by equation (7.21) the band edge energy shift becomes:

$$\delta E_c^{\text{cubic}} = 2a_c \left(1 - \frac{C_{12}}{C_{11}}\right) \epsilon_1 \tag{7.52}$$

where ϵ_1 is the in-plane strain due to lattice mismatch with the substrate and the C_{ij} are the elastic stiffness constants as previously defined. Similarly, for the hexagonal crystal geometry in the $[0001]$ direction the energy shift is given as:

$$\delta E_c^{\text{hex}} = 2 \left(a_{c\perp} - a_{c\parallel} \frac{C_{13}}{C_{33}}\right) \epsilon_1 \tag{7.53}$$

The conduction-band shift depends significantly on the material deformation potential constants and the lattice mismatch, and can be up to a few hundreds of millielectronvolts as shown in Fig. 7.6. It is interesting to see that, although the wurtzite AlGaN layer grown on GaN is highly strained in the growth plane, because of the anisotropy of the hexagonal crystals the overall shift is smaller.

Consider now the influence of the conduction-band edge shift on the electronic levels in a single quantum well. For simplicity it is assumed that the well material is strained while the barriers are matched to the substrate (Fig. 7.7). The conduction-band edge shift implies the equivalent change in the barrier height, defining the new quantum well potential as:

$$U(z)_{\text{str}} = U(z) + \delta E_c(z) \tag{7.54}$$

which results in a shift of the discrete energy levels localised within the quantum well.

In order to illustrate the importance of strain in quantum wells, Fig. 7.8 gives the energy of the ground state (with respect to the quantum well band edge) with and without the strain perturbation, as a function of the alloy composition (x) in the InP/$In_xAs_{1-x}P$ quantum well strained on an InP substrate. The calculation shows an average energy shift of around 10%, which can significantly influence the designs and applications. In addition, strain will influence the energy difference between subbands, so introducing a shift in intersubband transition energies. This implies that strain 'engineering' could be used to reach desired emission and detection wavelengths otherwise unattainable (see, for example, Faist *et al.* [7]).

Finally, it is worth noting that these calculations refer only to the Γ point in the conduction band of a crystal. In indirect bandgap semiconductors, the lowest-energy points in the conduction band lie far from Γ, and additional factors must be considered. Although a

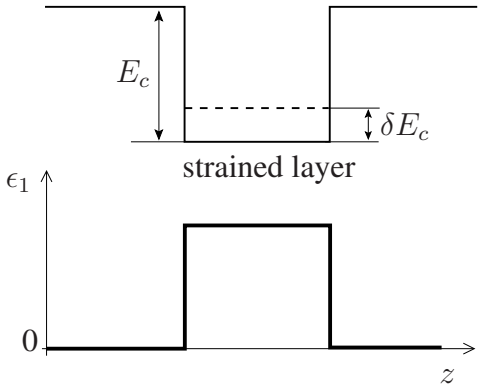


Figure 7.7: Schematic diagram of the conduction-band edge of a quantum well with a strain-induced energy band shift (δE_c). The original conduction-band discontinuity is E_c. The bottom diagram shows the corresponding in-plane strain distribution across the quantum well

detailed discussion of these effects is beyond the scope of this book, they are well documented in the literature (see [8, 9]) and a qualitative understanding may be gained by considering the symmetry of the system. For example, in bulk silicon, the conduction-band minima lie in six equivalent 'valleys' close to each of the 'X' symmetry points. When a biaxial stress is applied, the crystal is deformed such that the lattice constant is different in the growth direction from that in-plane with the substrate. Therefore, the symmetry of the crystal is broken, and the two conduction-band valleys aligned in the growth direction are shifted to a different energy from the four in-plane valleys. This 'uniaxial deformation' effect has significant implications for the design of devices in indirect bandgap materials, since inter-valley scattering processes can lead to leakage currents or charge trapping effects [10].

7.6 The piezoelectric effect

Certain types of crystal materials exhibit a behaviour such that, under stress, an extra electric charge gathers on their surfaces. The effect is called 'piezoelectricity' and is a consequence of a non-compensated electric polarisation generated in the volume of the crystal. Piezoelectric behaviour is exhibited in quartz, Rochelle salt and tourmaline, as well as in cubic and hexagonal semiconductors such as InGaAs and GaN. The changes in the crystal lattice geometry due to stress act to separate the charges and break the local electrostatic neutrality in the crystal; a schematic diagram illustrating this effect for wurtzite GaN is given in Fig. 7.9. The uncompensated positive and negative charges induce a global polarisation vector, the properties of which are defined by the material and the nature of the applied stress.

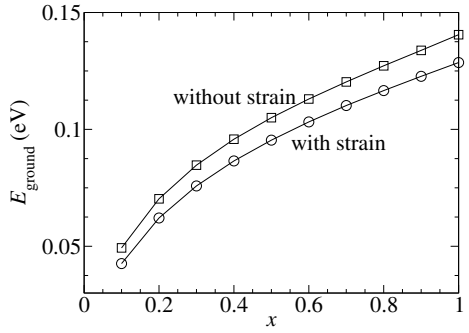


Figure 7.8: Energy of the ground state (with respect to the conduction-band edge in the well) in the InP/$In_xAs_{1-x}P$ system with and without strain, as a function of the alloy composition (x). The well width is set to 50 Å and the barrier width to 60 Å

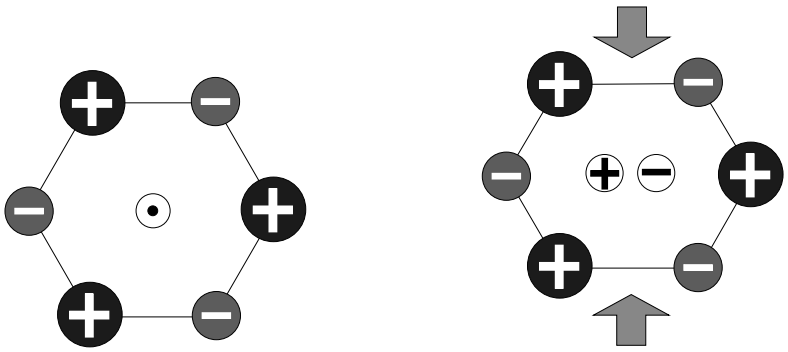


Figure 7.9: Schematic illustration of the charge separation in GaN grown along the $[0001]$ direction which causes the appearance of an uncompensated polarisation: (a) unstrained crystal lattice; (b) atomic positions when under strain

If an arbitrary stress is acting on the crystal, then the total induced polarisation (in matrix notation similar to that in Section 7.1) can be written as:

$$P_i = \sum_{k=1}^{6} d_{ik}\sigma_k \tag{7.55}$$

where d_{ik} is the piezoelectric modulus. The piezoelectric polarisation is a vector which is described by three components, P_1 and P_2 in the growth plane and P_3 along the growth direction, which in turn define a 3×6 matrix of the piezoelectric moduli (d). As for the case of the elastic constants, the symmetry in the crystals reduces the number of independent

elements of the d matrix. For cubic and hexagonal crystal geometries, the number of independent moduli relies on the crystal class, e.g. one modulus for InGaAs and three moduli for wurtzite GaN.

So it is stress-induced symmetry breaking in a crystal that leads to a piezoelectric polarisation, hence the underlying symmetry of the crystal is important. A good example of this is InGaAs. When grown on a GaAs substrate in the $[111]$ direction it exhibits a piezoelectric behaviour. However, layers grown along the $[001]$ direction are polarisation-free, while those grown along the $[110]$ direction exhibit a piezoelectric polarisation in the plane of the layers.

In the following, the piezoelectric properties of epitaxial layers will be considered in more detail for the example of the increasingly widespread wurtzite GaN. In order to determine the piezoelectric polarisation for GaN grown in the [0001] direction, consider equation (7.55) which, assuming a biaxial stress, can be written as:

$$\begin{bmatrix} P_1 \\ P_2 \\ P_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & \frac{1}{2}d_{15} & 0 \\ 0 & 0 & 0 & \frac{1}{2}d_{15} & 0 & 0 \\ d_{31} & d_{31} & d_{33} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{7.56}$$

Note that the factor $\frac{1}{2}$ is a consequence of adopting the matrix notation as discussed in Section 7.1. The matrix equation can be decoupled into three linear equations defining the polarisation vector components, the solution of which gives a single non-zero component of the piezoelectric polarisation in the growth direction as:

$$P_3 = d_{31}(\sigma_1 + \sigma_2) = 2d_{31}\sigma_1 \tag{7.57}$$

Using equations (7.20) and (7.21), the stress component σ_1 can be expressed in terms of the strain as $\sigma_1 = \epsilon_1 \left(C_{11} + C_{12} - 2\frac{C_{13}^2}{C_{33}} \right)$. Hence, the final form of the in-plane piezoelectric polarisation of GaN reads:

$$P_3 = 2d_{31}\epsilon_1 \left(C_{11} + C_{12} - 2\frac{C_{13}^2}{C_{33}} \right) \tag{7.58}$$

where the C_{ij} are the elastic stiffness constants. As the term $C_{11} + C_{12} - 2C_{13}^2/C_{33} > 0$ is always greater than zero, the sign of the polarisation is determined by the sign of the strain. Hence, P_3 is always positive for layers under biaxial compressive strain and negative for layers under tensile strain.

A more often used approach for describing piezoelectric properties of crystals than the moduli is that of the piezoelectric constants defined as:

$$e_{kl} = \sum_{j=1}^{6} d_{kj} C_{jl} \tag{7.59}$$

where the C_{jl} are the elastic stiffness constants. Then equation (7.55) for the piezoelectric polarisation can be rewritten as:

$$P_i = \sum_{k=1}^{6} e_{ik}\epsilon_k \tag{7.60}$$

In terms of these piezoelectric constants, the polarisation in the wurtzite GaN layer under biaxial strain now reads:

$$P_3 = \epsilon_1 e_{31} + \epsilon_2 e_{32} + \epsilon_3 e_{33} = 2\epsilon_1 e_{31} + \epsilon_3 e_{33} = 2\epsilon_1 \left(e_{31} - e_{33}\frac{C_{13}}{C_{33}} \right) \tag{7.61}$$

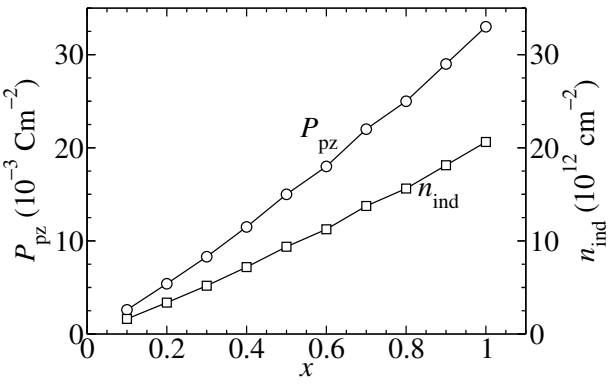


Figure 7.10: The piezoelectric polarisation and induced charge at the interface in a single $Al_xGa_{1-x}N$ layer grown on a GaN substrate as a function of the layer composition x

Figure 7.10 shows the calculated piezoelectric polarisation and the induced charge ($n_{\text{ind}} = P_{\text{pz}}/q$) at the growth interface for a single AlGaN layer grown strained on a GaN substrate. For example, for 50% of Al in the epitaxial layer the sheet electron density is around $1 \times 10^{13}\,\text{cm}^{-2}$. This is a common case in GaN-based HEMTs where the induced charge can significantly influence the electronic properties of devices. Furthermore, in quantum well structures the piezoelectric polarisation can have a significant effect on the band profile through the electric fields induced.

7.7 Induced piezoelectric fields in quantum wells

Consider a single quantum well structure based on piezoelectric active material. The different piezoelectric and strain properties of the well and the barrier materials will result in different polarisations. Hence, at the interfaces a gradient of the piezoelectric polarisation appears, which induces a fixed charge which gathers in the vicinity of the interface. The charge density is given by the equation:

$$\rho_p = -\nabla \mathbf{P} \tag{7.62}$$

Constant strain across the epitaxial layer implies a constant polarisation as well (see equation (7.61)). Hence, the change of the piezoelectric polarisation is abrupt and the induced charge can be found as the difference between the polarisations in the adjacent layers. Furthermore, this accumulated charge at the interfaces induces an electric field in order to satisfy Gauss's law.

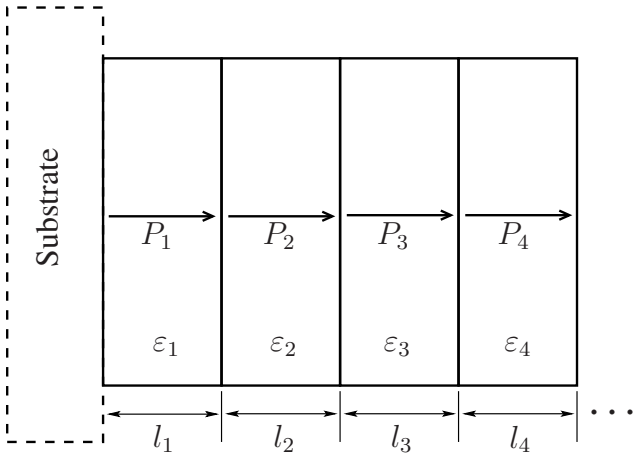


Figure 7.11: Schematic diagram of a multilayer structure based on piezoelectric material, giving the notation used in text

Consider now the general case of the multilayer structure in Fig. 7.11 in order to estimate the intrinsic electric fields in the layers. The electrostatic boundary conditions of the displacement vector ($\mathbf{D}$) at the adjacent interfaces are applied as:

$$\mathbf{D}_i = \mathbf{D}_{i+1} \tag{7.63}$$

where linear material properties[3] are assumed, i.e. $\mathbf{D}_i = \varepsilon_i \mathbf{E}_i + \mathbf{P}_i$. Though the piezoelectric polarisation vector can have an arbitrary direction only the component along the growth direction will influence the electronic properties of a quantum well structure and therefore only this component will be considered here. For an arbitrary n-layer structure,

[3] ε is used to represent the permittivity in this chapter to distinguish it from ϵ, which is the commonly accepted symbol for the components of strain.

equation (7.63) gives a system of $n-1$ linear equations:

$$\begin{aligned}
\varepsilon_1 F_1 + P_1 &= \varepsilon_2 F_2 + P_2 \\
\varepsilon_2 F_2 + P_2 &= \varepsilon_3 F_2 + P_3 \\
&\vdots \\
\varepsilon_{i-1} F_{i-1} + P_{i-1} &= \varepsilon_i F_i + P_i \\
&\vdots \\
\varepsilon_{n-1} F_{n-1} + P_{n-1} &= \varepsilon_n F_n + P_n
\end{aligned} \tag{7.64}$$

which can be solved provided appropriate boundary conditions across the structure are defined. Though there are no constraints on the choice of the boundary condition, the natural choice are the 'hard wall' boundary conditions, which set the total potential drop across the structure to zero. All other possibilities would imply that the quantum well structure acts as a voltage source in the closed circuit. For a periodic structure (e.g. a superlattice) the hard wall and the periodic boundary conditions are equivalent.

Knowing the boundary condition, the necessary nth equation of the system is defined and reads:

$$\sum_{i=1}^{n} F_i l_i = 0 \tag{7.65}$$

where L_i is the width of the ith layer.

The solution of the system gives a rather simple equation:

$$F_j = \frac{\sum_k (P_k - P_j)\,\frac{l_k}{\varepsilon_k}}{\varepsilon_j \sum_k \frac{l_k}{\varepsilon_k}} \tag{7.66}$$

from which the intrinsic electric field can be easily calculated. It is interesting to see that in the quantum wells the electric field exists even in the layers with zero piezoelectric polarisation; this is a consequence of the charge induced at the interfaces with the adjacent piezoelectric layers.

In the case of wurtzite GaN-based quantum wells, the piezoelectric field due to the large strain and piezoelectric constants can be stronger than $1\,\mathrm{MV\,cm^{-1}}$ [11]. Figure 7.12 shows such an example, where the piezoelectric field in the well and barrier layers of a single GaN/AlGaN quantum well structure grown on a GaN substrate is calculated as a function of the barrier layer composition. The electric field in an unstrained 20 Å wide layer is as large as $5\,\mathrm{MV\,cm^{-1}}$, while in a thicker 600 Å strained AlGaN layer it is proportionally smaller but still around $1\,\mathrm{MV\,cm^{-1}}$. The piezoelectric field in GaN alloys is an order of magnitude larger than in another widely used material, cubic InGaAs, which exhibits piezoelectric fields of $\sim 100\,\mathrm{kV\,cm^{-1}}$ [12].

7.8 Effect of piezoelectric fields on quantum wells

As shown in the previous section, the piezoelectric effect manifests itself as an electric field in the layers of the quantum well structure. To explore the influence of such fields on the

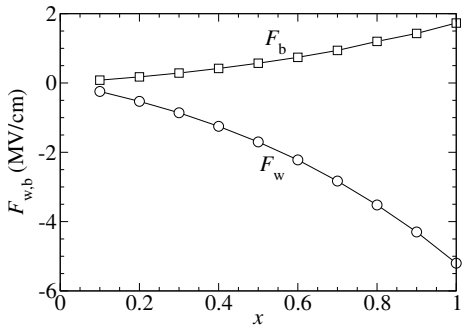


Figure 7.12: The induced piezoelectric field in a GaN well and $Al_xGa_{1-x}N$ barrier layers grown on a GaN substrate as a function of the barrier layer composition x

band profile and the electronic structure, the simplest case of a single quantum well of width l_w, surrounded by a pair of barriers of width l_b, will be considered. Also, without loss of generality, a uniform dielectric constant can be assumed across the structure. Hence, the magnitude of the piezoelectric fields in the quantum well, assuming identical barriers, follows from equation (7.66) as:

$$F_\mathrm{w} = \frac{2\,(P_\mathrm{b} - P_\mathrm{w})\,l_\mathrm{b}}{\varepsilon(l_\mathrm{w} + l_\mathrm{b})} \tag{7.67}$$

and

$$F_\mathrm{b} = \frac{(P_\mathrm{w} - P_\mathrm{b})\,l_\mathrm{w}}{\varepsilon(l_\mathrm{w} + l_\mathrm{b})} \tag{7.68}$$

where P_w and P_b are the polarisation magnitudes in the well and the barrier respectively, and ε is the constant permittivity. Note that the factor of 2 in the first equation arises from the contributions of the two identical barrier layers.

The electric fields in the well and the barrier layer must be of opposite sign, as the overall potential drop across the quantum well has to be zero. This can be expressed analytically using equation (7.65) as:

$$F_\mathrm{w} l_\mathrm{w} + 2F_\mathrm{b} l_\mathrm{b} = 0 \tag{7.69}$$

It is straightforward to confirm that the fields derived above satisfy this condition. This implies that the distribution of the electric field across the structure is proportional to the ratio of the layer widths. Consequently, the thinner epitaxial layer will have a higher electric field than that of the thicker.

The piezoelectric field alters the quantum well band profile, which can be represented as a step-linear potential (with respect to the coordinate system as in Fig. 7.13):

$$V(z) = \begin{cases} eF_\mathrm{b} z + V_\mathrm{b}, & z < 0 \\ eF_\mathrm{w} z, & 0 < z < l_\mathrm{w} \\ eF_\mathrm{b} z + F_\mathrm{w} l_\mathrm{w} + V_\mathrm{b}, & z > l_\mathrm{w} \end{cases} \tag{7.70}$$

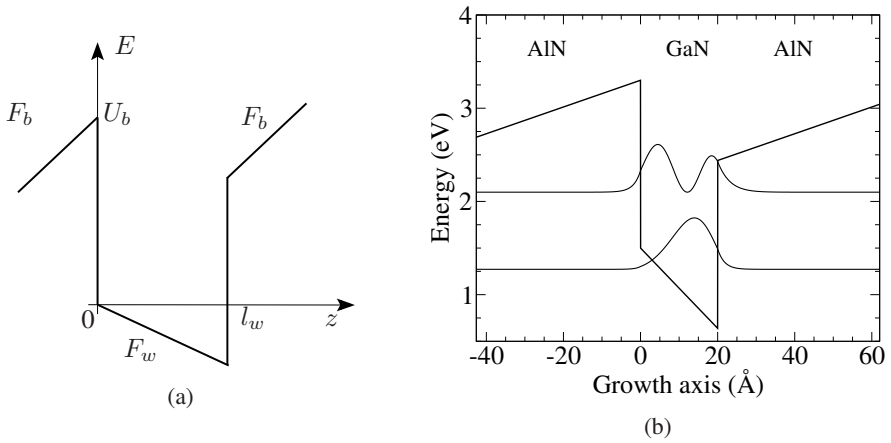


Figure 7.13: (a) Schematic diagram of typical induced electric fields in a quantum well based on piezoelectric materials and (b) the band profile of a GaN/AlN quantum well structure with wave function moduli of the first two bound states

where l_w is the well width and V_b is the barrier height. The previous potential definition is equivalent to a biased quantum well structure and the solutions can be found as a linear combination of Airy's functions as discussed in Chapter 2.

Another important effect is that the piezoelectric field can be screened by free electrons. The linear potential causes electrons to gather close to the interfaces, therefore inducing space charges. If the carrier density in the quantum well is comparable to the density of the piezoelectric induced charges, the electrostatic field due to the free electrons can suppress the piezoelectric field. However, as the induced charge at the interfaces is usually much larger than the quantum well doping limit, this screening effect is generally not pronounced in applications.

The change of the band profile due to a piezoelectric field can have a strong influence on the energies of intersubband transitions. This is illustrated in Fig. 7.14(a), which shows a comparison of the transition energies between the lowest two subbands in GaN/AlGaN quantum wells with and without the internal electric field included. A substantial increase (blue shift) of the intersubband transition energies is clearly observed and has been observed experimentally [13].

The piezoelectric field can also considerably modify the localisation properties of the wave functions, increasing or decreasing the overlap between the ground and other excited states. This can reflect on the dipole matrix element (defined as $z_{ij} = \langle \Psi_i | z | \Psi_j \rangle$) and consequently on the intersubband optical properties of quantum well structures such as the intersubband absorption ($A_{ij} \sim |z_{ij}|^2$). For GaN/AlGaN quantum wells, the relative change in the dipole matrix element z_{12} is shown in Fig. 7.14(b). Clearly, the estimated decrease of up to 50% in z_{ij} is enough to significantly deteriorate the performance of a device in a potential application. A similar problem is present in quantum well lasers based on interband

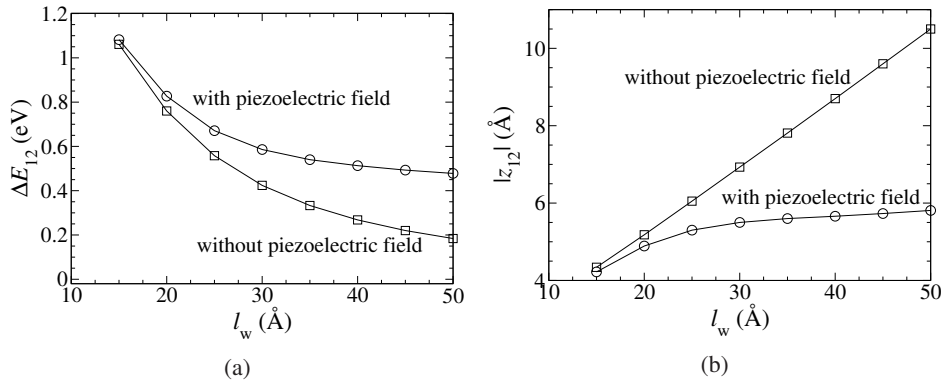


Figure 7.14: Illustration of the influence of the piezoelectric field on the electronic and optical properties of GaN/AlN quantum wells, (a) the energy difference and (b) the dipole matrix element, between the first two bound states with and without the piezoelectric field

transitions between the conduction and the valence bands (see [14]).

The previously mentioned effects on optical transition energies and absorption strengths highlight the importance of a thorough understanding of the piezoelectric field properties for the modelling and design of optoelectronic devices.

Exercises

All necessary material parameters are given in Appendix A.

(1). Consider a film of InAs grown epitaxially upon a GaAs substrate.

- (a) Calculate the stress and strain tensors.
- (b) Find the volume dilation.
- (c) Calculate the elastic energy density in the film.
- (d) Determine the critical thickness for growth of the material.
- (e) Hence, estimate the total elastic energy stored within a critically thick film of InAs grown on a 2-inch GaAs wafer.

(2). A superlattice consisting of $100\,\text{Å}$ GaAs wells and $50\,\text{Å}$ $In_{0.25}Ga_{0.75}As$ barriers is grown upon an $In_xGa_{1-x}As$ buffer layer.

- (a) Find the optimal alloy composition of the buffer layer, in order to strain-balance the structure. Assume identical elastic constants in each material.
- (b) Repeat the calculation, but take into account the effect of composition-dependent elastic constants. Comment on the difference in result.
- (c) Assuming that the buffer composition is fixed at $x = 0.1$, determine a new different barrier thickness that would satisfy the strain-balancing condition. For

simplicity, ignore the spatial variation in elastic constant. Repeat this calculation, using barriers with the original thickness but a different alloy composition. Discuss the implications of this study for the design of practical heterostructure devices.

(3). Consider a double quantum well structure, consisting of a pair of identical wells of width l_{w}, separated by a barrier of width l_0 and surrounded by a pair of identical barriers of width l_{b}.

(a) Assuming a constant dielectric constant, determine an expression for the piezoelectric field in each layer of the structure.

(b) Confirm that the overall potential drop across the structure is zero.

(c) Determine the difference in potential between the centres of the two wells in terms of the widths and the polarisation in each layer.

(d) Confirm that the potential drop converges to $\Delta V = eF_{\mathrm{w}}l_{\mathrm{w}}$ as the central barrier width reduces to zero.

References

[1] J. F. Nye, *Physical Properties of Crystals*, Clarendon Press, Oxford, 1957.

[2] J. W. Matthews and A. E. Blakeslee, 'Defects in epitaxial multilayers: I. Misfit dislocations', *J. Crystal Growth*, **27**:118–125, 1974.

[3] A. Braun, K. M. Briggs, and P. Böni, 'Analytical solution to Matthews' and Blakeslee's critical dislocation formation thickness of epitaxially grown thin films', *J. Crystal Growth*, **241**(1–2):231–234, 2002.

[4] M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, and F. Rossi, *GNU Scientific Library Reference Manual.*, Network Theory Limited, Bristol, Third (v1.12) edition, 2009.

[5] D. J. Paul, 'Si/SiGe heterostructures: from material and physics to devices and circuits', *Semicond. Sci. Technol.*, **19**(10):R75, 2004.

[6] V. D. Jovanović, Z. Ikonić, D. Indjin, P. Harrison, V. Milanović, and R. A. Soref, 'Designing strain-balanced GaN/AlGaN quantum well structures: Application to intersubband devices at 1.3 and 1.55 μm', *J. Appl. Phys.*, **93**:3194, 2003.

[7] J. Faist, F. Capasso, D. L. Sivco, A. L. Hutchinson, S.-N. G. Chu, and A. Y. Cho, 'Short wavelength (λ˜3.4 μm) quantum cascade laser based on strained compensated InGaAs/AlInAs', *Appl. Phys. Lett.*, **72**(6):680–682, 1998.

[8] M. M. Rieger and P. Vogl, 'Electronic-band parameters in strained $Si_{1-x}Ge_x$ alloys on $Si_{1-y}Ge_y$ substrates', *Phys. Rev. B*, **48**(19):14276–14287, 1993.

[9] S. Smirnov and H. Kosina, 'Monte Carlo modeling of the electron mobility in strained $Si_{1-x}Ge_x$ layers on arbitrarily oriented $Si_{1-y}Ge_y$ substrates', *Solid-State Electron.*, **48**(8):1325–1335, 2004.

[10] A. Valavanis, T. V. Dinh, L. J. M. Lever, Z. Ikonić, and R. W. Kelsall, 'Material configurations for n-type silicon-based terahertz quantum cascade lasers', *Phys. Rev. B*, **83**(19):195321, 2011.

[11] O. Ambacher, J. Majewski, C. Miskys, A. Link, M. Hermann, M. Eickhoff, M. Stutzmann, F. Bernardini, V. Fiorentini, V. Tilak, B. Schaff, and L. F. Eastman, 'Pyroelectric properties of Al(In)GaN/GaN hetero- and quantum well structures', *J. Phys.: Condens. Matter*, **14**:3399, 2002.

[12] D. L. Smith and C. Mailhiot, 'Theory of semiconductor superlattice electronic structure', *Rev. Mod. Phys.*, **62**:173, 1990.

[13] C. Gmachl, H. M. Ng, and A. Y. Cho, 'Intersubband absorption in GaN/AlGaN multiple quantum wells in the wavelength range of $\lambda \sim$1.75–4.2 μm', *Appl. Phys. Lett.*, **77**:334, 2000.

[14] J. S. Im, H. Kollmer, J. Off, A. Sohmer, F. Scholz, and A. Hangleiter, 'Reduction of oscillator strength due to piezoelectric fields in GaN/Al_xGa_{1-x}N quantum wells', *Phys. Rev. B*, **57**(16):R9435–R9438, 1998.

8

Simple models of quantum wires and dots

8.1 Further confinement

It has already been shown that the reduction in dimensionality produced by confining electrons (or holes) to a thin semiconductor layer leads to a dramatic change in their behaviour. This principle can be developed by further reducing the dimensionality of the electron's environment from a two-dimensional quantum well to a one-dimensional quantum wire and eventually to a zero-dimensional quantum dot. In this context, of course, the dimensionality refers to the number of degrees of freedom in the electron momentum: in fact, within a quantum wire, the electron is confined across two directions, rather than just the one in a quantum well, thus reducing the degrees of freedom to one; in a quantum dot, the electron is confined in all three dimensions, thus reducing the degrees of freedom to zero. If the number of degrees of freedom is denoted by $\mathcal{D}_\mathrm{f}$ and the number of directions of confinement is denoted by $\mathcal{D}_\mathrm{c}$, then clearly:

$$\mathcal{D}_\mathrm{f} + \mathcal{D}_\mathrm{c} = 3 \tag{8.1}$$

for all solid state systems. These values are highlighted for the four possibilities shown in Table 8.1. Tradition has determined that the reduced-dimensionality systems are labelled by the remaining degrees of freedom in the electron motion, i.e. $\mathcal{D}_\mathrm{f}$, rather than the number of directions with confinement $\mathcal{D}_\mathrm{c}$.

Table 8.1 The number of degrees of freedom $\mathcal{D}_\mathrm{f}$ in the electron motion, together with the extent of the confinement $\mathcal{D}_\mathrm{c}$, for the four basic dimensionality systems

System	Directions of confinement, $\mathcal{D}_\mathrm{c}$	Degrees of freedom, $\mathcal{D}_\mathrm{f}$
Bulk	0	3
Quantum well	1	2
Quantum wire	2	1
Quantum dot	3	0

Figure 8.1 gives a simple outline of how quantum wires *might* be fabricated, although note that there is more than one method and the interested reader should refer to a specialist

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

growth treatise for further details. A standard quantum well layer can be patterned with photolithography or perhaps electron-beam lithography, and etched to leave a free-standing strip of quantum well material; the latter may or may not be filled in with an overgrowth of the barrier material (in this case, $Ga_{1-x}Al_xAs$). Any charge carriers are still confined along the heterostructure growth (z-)axis, as they were in the quantum well, but in addition (provided the strip is narrow enough) they are now confined along an additional direction, either the x- or the y-axis, depending on the lithography.

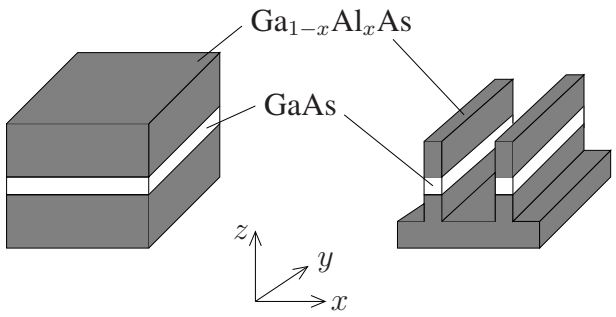


Figure 8.1: Fabrication of quantum wires

Figure 8.2 shows an expanded view of a single quantum wire, where clearly the electron (or hole) is free to move in only one direction, in this case along the y-axis. Within the effective mass approximation the motion along the axis of the wire can still be described by a parabolic dispersion, i.e.

$$E = \frac{\hbar^2 k^2}{2m^*} \tag{8.2}$$

just as in bulk and for the in-plane motion within a quantum well.

Another class of quantum wire can be formed by patterning the substrate *before* growth. This leads to the formation of so-called *V-grooved quantum wires*, (see [1], p. 35); the solution of these has been dealt with by Gangopadhyay and Nag [2].

Quantum dots (see [3] for an introduction to their applications) *can* again be formed by further lithography and etching; e.g. if a quantum well sample is etched to leave pillars rather than wires, then a charge carrier can become confined in all three dimensions, as illustrated in Fig. 8.3.

Under certain growth conditions, when a thin layer of a semiconductor is grown on top of a substrate which has a quite different lattice constant, then in an attempt to minimise the total strain energy between the bonds, the thin layer spontaneously orders or *self-assembles* into quantum dots. Microscopy has shown the dots to take the shape of pyramids, or square based 'tetrahedra' [4, 5].

In the following sections example solutions will be derived for some of the simpler geometries; some of the more complex ones will have to wait until Chapter 16.

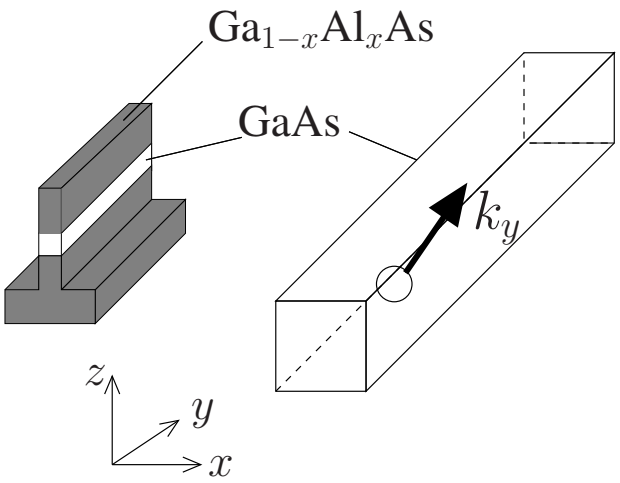


Figure 8.2: A single wire and an expanded view showing schematically the single degree of freedom in the electron momentum

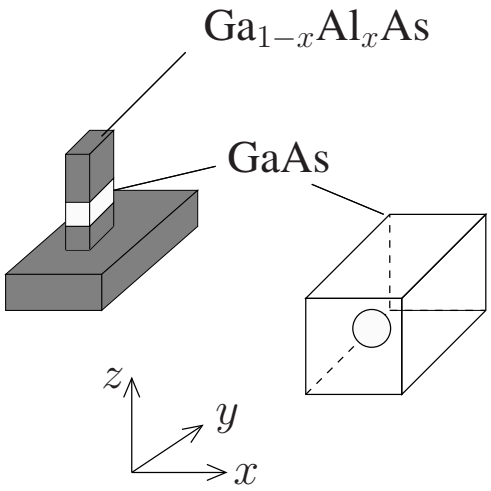


Figure 8.3: A single free-standing pillar containing a quantum dot and an expanded view which shows schematically the removal of all degrees of freedom for the electron momentum

8.2 Schrödinger's equation in quantum wires

The general three-dimensional Schrödinger equation for constant effective mass is:

$$-\frac{\hbar^2}{2m^*}\nabla^2\psi(x,y,z) + V(x,y,z)\psi(x,y,z) = E\psi(x,y,z) \tag{8.3}$$

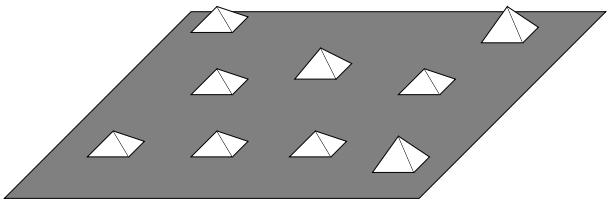


Figure 8.4: Schematic representation of the pyramidal shape of self-assembled quantum dots in highly lattice-mismatched systems

In analogy to the in-plane dispersion discussed in Section 2.2, in a quantum wire it is possible to decouple the motion along the length of the wire. Taking the axis of the wire along x, the total potential $V(x, y, z)$ can always be written as the sum of a two-dimensional confinement potential plus the potential along the wire (which happens to be zero in this case), i.e.

$$V(x, y, z) = V(x) + V(y, z) \tag{8.4}$$

The eigenfunction can then be written as a product of two components:

$$\psi(x, y, z) = \psi(x)\psi(y, z) \tag{8.5}$$

Substituting both equations (8.4) and (8.5) into equation (8.3):

$$-\frac{\hbar^2}{2m^*}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\psi(x)\psi(y, z) + (V(x) + V(y, z))\,\psi(x)\psi(y, z) = E\psi(x)\psi(y, z) \tag{8.6}$$

Writing the energy as a sum of terms associated with the two components of the motion:

$$-\frac{\hbar^2}{2m^*}\left[\psi(y, z)\frac{\partial^2\psi(x)}{\partial x^2} + \psi(x)\frac{\partial^2\psi(y, z)}{\partial y^2} + \psi(x)\frac{\partial^2\psi(y, z)}{\partial z^2}\right] + \psi(y, z)V(x)\psi(x) + \psi(x)V(y, z)\psi(y, z) = (E_x + E_{y,z})\psi(x)\psi(y, z) \tag{8.7}$$

It is now possible to associate distinct kinetic and potential energies on the left-hand side of equation (8.7) with the components E_x and $E_{y,z}$ on the right-hand side, thus giving two *decoupled* equations, as follows:

$$-\frac{\hbar^2}{2m^*}\psi(y, z)\frac{\partial^2\psi(x)}{\partial x^2} + \psi(y, z)V(x)\psi(x) = \psi(y, z)E_x\psi(x) \tag{8.8}$$

$$-\frac{\hbar^2}{2m^*}\left[\psi(x)\frac{\partial^2\psi(y, z)}{\partial y^2} + \psi(x)\frac{\partial^2\psi(y, z)}{\partial z^2}\right] + \psi(x)V(y, z)\psi(y, z) = \psi(x)E_{y,z}\psi(y, z) \tag{8.9}$$

$\psi(y,z)$ is not acted upon by any operator in (8.8), and similarly for $\psi(x)$ in (8.9), and thus they can be divided out. In addition, as mentioned above, the potential component along the axis of the wire $V(x) = 0$, thus giving the final decoupled equations of motion as follows:

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2\psi(x)}{\partial x^2} = E_x\psi(x) \tag{8.10}$$

$$-\frac{\hbar^2}{2m^*}\left[\frac{\partial^2\psi(y,z)}{\partial y^2} + \frac{\partial^2\psi(y,z)}{\partial z^2}\right] + V(y,z)\psi(y,z) = E_{y,z}\psi(y,z) \tag{8.11}$$

Clearly, the first of these equations (equation (8.10)) is satisfied by a plane wave of the form $\exp(ik_x x)$, thus giving the standard dispersion relationship:

$$E_x = \frac{\hbar^2 k_x^2}{2m^*} \tag{8.12}$$

The second of these equations of motion (equation (8.11)) is merely the Schrödinger equation for the two-dimensional confinement potential characterising a quantum wire. For a general cross-sectional wire, equation (8.11) should really be solved by using a full two-dimensional solution. For now though, special cases of the solution of equation (8.11), for the relevant commonly found geometries, will be illustrated. Such particular solutions rely upon the ability to further decouple the motion into independent components.

8.3 Infinitely deep rectangular wires

Perhaps the simplest quantum wire geometry would be a rectangular cross-section surrounded by infinite barriers. This is illustrated schematically in Fig. 8.5 and can be considered to be the two-dimensional analogy to the one-dimensional confinement potential of the standard infinitely deep quantum well.

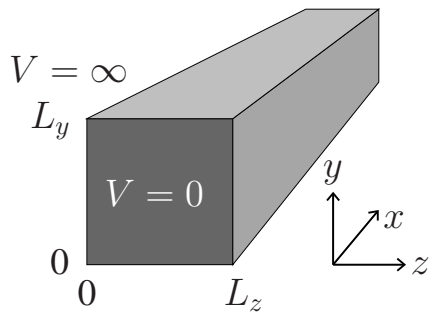


Figure 8.5: The infinitely deep rectangular cross-section quantum wire

Within the quantum wire, the potential is zero, while outside the wire it is infinite; thus in the latter case the wave function is zero. Hence, the Schrödinger equation is only defined

within the wire for the motion in the two confined y- and z-directions, i.e. equation (8.11) becomes:

$$-\frac{\hbar^2}{2m^*}\left[\frac{\partial^2\psi(y,z)}{\partial y^2}+\frac{\partial^2\psi(y,z)}{\partial z^2}\right]=E_{y,z}\psi(y,z) \tag{8.13}$$

The form of the potential in this Schrödinger equation (i.e. zero!) allows the motions to be decoupled further, writing:

$$\psi(y,z)=\psi(y)\psi(z) \tag{8.14}$$

and then:

$$-\frac{\hbar^2}{2m^*}\psi(z)\frac{\partial^2\psi(y)}{\partial y^2}-\frac{\hbar^2}{2m^*}\psi(y)\frac{\partial^2\psi(z)}{\partial z^2}=E_{y,z}\psi(y)\psi(z) \tag{8.15}$$

Again, it is possible to associate the individual kinetic energy terms on the left-hand side of equation (8.15) with separate energy components, i.e. by writing $E_{y,z}=E_y+E_z$, then:

$$-\frac{\hbar^2}{2m^*}\psi(z)\frac{\partial^2\psi(y)}{\partial y^2}-\frac{\hbar^2}{2m^*}\psi(y)\frac{\partial^2\psi(z)}{\partial z^2}=\psi(z)E_y\psi(y)+\psi(y)E_z\psi(z) \tag{8.16}$$

The decoupling is completed with the following:

$$-\frac{\hbar^2}{2m^*}\psi(z)\frac{\partial^2\psi(y)}{\partial y^2}=\psi(z)E_y\psi(y) \tag{8.17}$$

$$-\frac{\hbar^2}{2m^*}\psi(y)\frac{\partial^2\psi(z)}{\partial z^2}=\psi(y)E_z\psi(z) \tag{8.18}$$

Dividing equation (8.17) by $\psi(z)$ and equation (8.18) by $\psi(y)$ gives:

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2\psi(y)}{\partial y^2}=E_y\psi(y) \tag{8.19}$$

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2\psi(z)}{\partial z^2}=E_z\psi(z) \tag{8.20}$$

Given that the potential outside the wire is infinite, then the standard boundary condition of continuity in the wave function implies that *both* $\psi(y)$ and $\psi(z)$ are zero at the edges of the wire. Thus, equations (8.19) and (8.20) are identical to those of the one-dimensional infinitely deep quantum well—the two-dimensional Schrödinger equation has been decoupled into two one-dimensional equations. Given the origin in a 'corner', and the wire dimensions L_y and L_z, as in Fig. 8.5, then the solutions follow as:

$$\psi(y)=\sqrt{\frac{2}{L_y}}\sin\left(\frac{\pi n_y y}{L_y}\right) \tag{8.21}$$

and

$$\psi(z)=\sqrt{\frac{2}{L_z}}\sin\left(\frac{\pi n_z z}{L_z}\right) \tag{8.22}$$

which give the components of energy as:

$$E_y=\frac{\hbar^2\pi^2 n_y^2}{2m^*L_y^2} \tag{8.23}$$

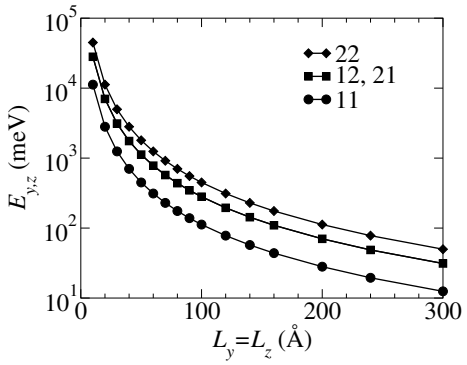


Figure 8.6: The confinement energy in an infinitely deep quantum wire with a square cross-section

and

$$E_z = \frac{\hbar^2 \pi^2 n_z^2}{2m^* L_z^2} \tag{8.24}$$

Thus, the total energy due to confinement, $E_{y,z} = E_y + E_z$, is:

$$E_{y,z} = \frac{\hbar^2 \pi^2}{2m^*} \left(\frac{n_y^2}{L_y^2} + \frac{n_z^2}{L_z^2} \right) \tag{8.25}$$

The confined states of a quantum wire are therefore described by the two principle quantum numbers n_y and n_z, in contrast to the single number required for the one-dimensional confinement potential in quantum wells.

Fig. 8.6 displays the confinement energies $E_{y,z}$ for (n_y,n_z) equal to $(1,1)$, $(1,2)$, $(2,1)$, and $(2,2)$ as a function of the side length $L_y = L_z$ for a *square* cross-section infinitely deep quantum wire. In this case of a square cross-sectional wire, the confinement energies of the $(1,2)$ and $(2,1)$ states are equal; however, clearly this will not be the case for a rectangular cross-section wire, which has $L_y \neq L_z$. Just as in the quantum wells met previously, the confinement energy decreases as the size of the system increases.

The wave function for a stationary (wave function along the length of the wire, independent of position) confined particle within the wire is real, and hence the charge density is simply $[\psi(y)\psi(z)]^2$. This is plotted in Fig. 8.7 for the four lowest confined states over a cross-sectional plane. The spatial distribution of the charge density is dependent upon the principal quantum numbers n_y and n_z (as would be expected), and the number of antinodes (local maxima) is equal to $n_y n_z$.

8.4 Simple approximation to a finite rectangular wire

More relevant to real devices would be a rectangular cross-sectional quantum wire, but with finite height barriers, as would be fabricated with a post-etch overgrowth. Figure 8.8 (left) illustrates the two-dimensional confinement potential $V(y,z)$ for this system. With this

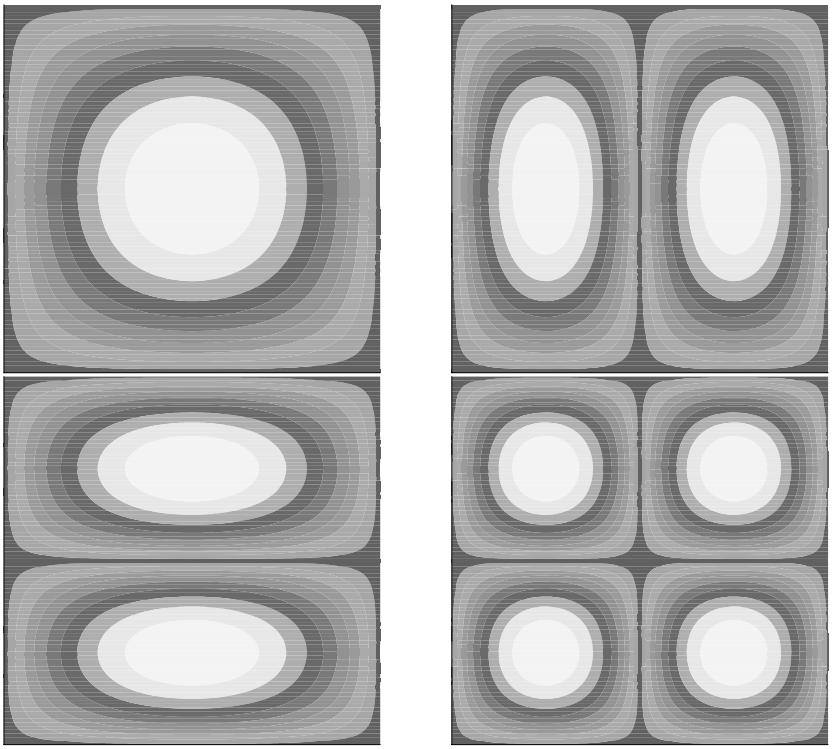


Figure 8.7: The charge densities of the four lowest-energy confined states within a infinitely deep quantum wire; (top left) n_y=1, n_z=1; (top right) n_y=1, n_z=2; (bottom left) n_y=2, n_z=1; (bottom right) n_y=2, n_z=2

configuration it is not possible to write the potential $V(y,z)$ as a sum of two independent potentials $V(y)$ and $V(z)$, and thus it is not possible to separate the y- and z-motions.

However, a very loose approximation may be to write the potential as in Fig. 8.8 (right). With this form, $V(y,z)$ does equal $V(y)+V(z)$, where $V(y)$ and $V(z)$ are independent finite well potentials, as in Section 2.7. The approximation occurs in the 'corner regions' outside of the wire where the two quantum well potential barrier heights V sum to give $2V$. This is in areas which are not expected to be sampled too much by the eigenfunctions, particularly those in wide wires and the lower-energy states.

Thus proceeding, the Schrödinger equation for the y–z cross-sectional confined motion in a general quantum wire is given by equation (8.11), i.e.

$$-\frac{\hbar^2}{2m^*}\left[\frac{\partial^2\psi(y,z)}{\partial y^2}+\frac{\partial^2\psi(y,z)}{\partial z^2}\right]+V(y,z)\psi(y,z)=E_{y,z}\psi(y,z) \qquad (8.26)$$

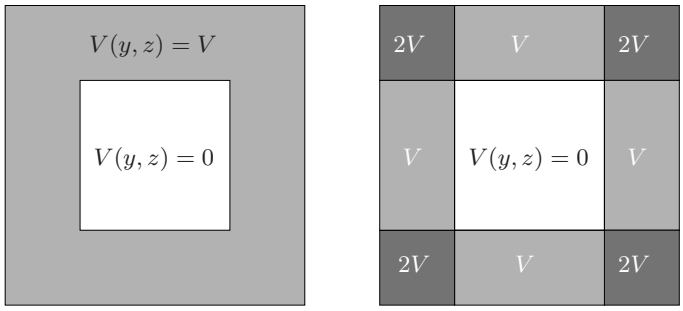


Figure 8.8: The rectangular cross-section quantum wire with finite barriers (left) and an approximate form for the potential (right), suitable for decoupling the motion

Therefore, writing the potential $V(y,z) = V(y) + V(z)$ and the wave function $\psi(y,z)$ as $\psi(y)\psi(z)$ gives:

$$-\frac{\hbar^2}{2m^*}\psi(z)\frac{\partial^2\psi(y)}{\partial y^2} - \frac{\hbar^2}{2m^*}\psi(y)\frac{\partial^2\psi(z)}{\partial z^2} + [V(y) + V(z)]\,\psi(y)\psi(z) = E_{y,z}\psi(y)\psi(z) \tag{8.27}$$

Again, by writing the energy $E_{y,z}$ as a sum of two components associated with the y- and z-motions, equation (8.27) can then be split into two, giving:

$$-\frac{\hbar^2}{2m^*}\psi(z)\frac{\partial^2\psi(y)}{\partial y^2} + \psi(z)V(y)\psi(y) = \psi(z)E_y\psi(y) \tag{8.28}$$

and

$$-\frac{\hbar^2}{2m^*}\psi(y)\frac{\partial^2\psi(z)}{\partial z^2} + \psi(y)V(z)\psi(z) = \psi(y)E_z\psi(z) \tag{8.29}$$

Dividing equation (8.28) by $\psi(z)$ and equation (8.29) by $\psi(y)$ gives familiar equations for straightforward one-dimensional potentials, i.e.

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2\psi(y)}{\partial y^2} + V(y)\psi(y) = E_y\psi(y) \tag{8.30}$$

and

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2\psi(z)}{\partial z^2} + V(z)\psi(z) = E_z\psi(z) \tag{8.31}$$

In this derivation, the kinetic energy operator appropriate for a constant effective mass has been employed, but equally well, as there is nothing that depends upon this form, the variable effective mass kinetic energy operator, $(\partial/\partial z)(1/m^*(z))(\partial/\partial z)$, could have been used. The final outcome is the same, namely that as the independent potentials $V(y)$ and $V(z)$ are simply those of a finite quantum well, the solutions for both the wave functions and the confinement energies follow as before (Section 2.7).

Figure 8.9 shows the equivalent of Fig. 8.6 but with finite barriers, corresponding to Al concentrations of 20% and 40% in $Ga_{1-x}Al_xAs$. The solutions were derived by combining

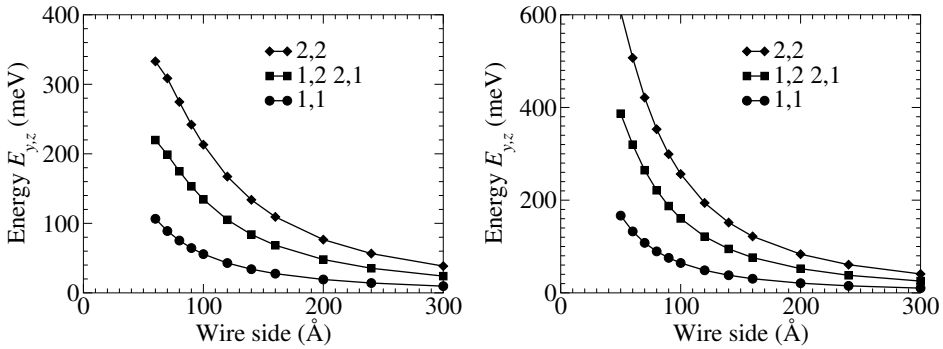


Figure 8.9: The confinement energy, $E_{y,z} = E_y + E_z$, in a square cross-section GaAs quantum wire surrounded by finite-height barriers of $Ga_{0.8}Al_{0.2}As$ (left) and $Ga_{0.6}Al_{0.4}As$ (right)

the eigenvalues from two independent one-dimensional quantum well calculations, including the effective mass mismatch at the well (in this case wire) interface. The eigenvalues labelled '1,1' correspond to the ground state of the y-motion combined with the ground state of the z-motion. The others represent the combinations formed between the lowest two energy states in each direction. The behaviour of the energy with the length of the wire side is qualitatively similar to the infinite-barrier case, i.e. there is degeneracy between the 1,2 and the 2,1 solutions, the ordering of the energy levels is the same and the eigenvalues decrease with increasing wire side.

The corresponding charge densities $[\psi(y,z)\psi(y,z)]$ are plotted in Fig. 8.10 for a cross-section of the wire for these same four eigenstates. Again, the behaviour is similar to the infinitely deep wire case, with the same distribution of maxima and minima. However, the main difference is, of course, that the finite barrier height allows for a significant 'leakage' of the wave function into the surrounding material. It is this interaction with the 'potential pillars' located outside of the wire at each corner, which originate from the approximated potential in Fig. 8.8, that limits the applicability or accuracy of this simple model.

For this relatively narrow ($100\,\text{Å} \times 100\,\text{Å}$) wire, the scale next to the ground state charge density in Fig. 8.10 (top left) implies that, roughly speaking, probably around 80–90% of the charge is confined within the wire. Thus the effect of the approximate potential outside of the wire will not be too substantial. However, for the higher-energy states more of the 'lighter shade' is outside the wire, and thus it would be expected that this approximation would be worse. The same effect would be true when reducing the length of the wire side.

This simple approach to the finite quantum wire is just a way of understanding the basic physics and being able to predict *qualitatively* how the electronic properties change with the system parameters. For a more complete description, a full two-dimensional Schrödinger solution may be required (see [6, 2]), or alternatively the empirical pseudo-potential approach (see Chapter 16) may be employed.

The energy eigenvalues $E_{y,z}$ *could* be improved by considering a perturbation on the two-dimensional system which removed the '$2V$' potential pillars. Using first-order perturbation

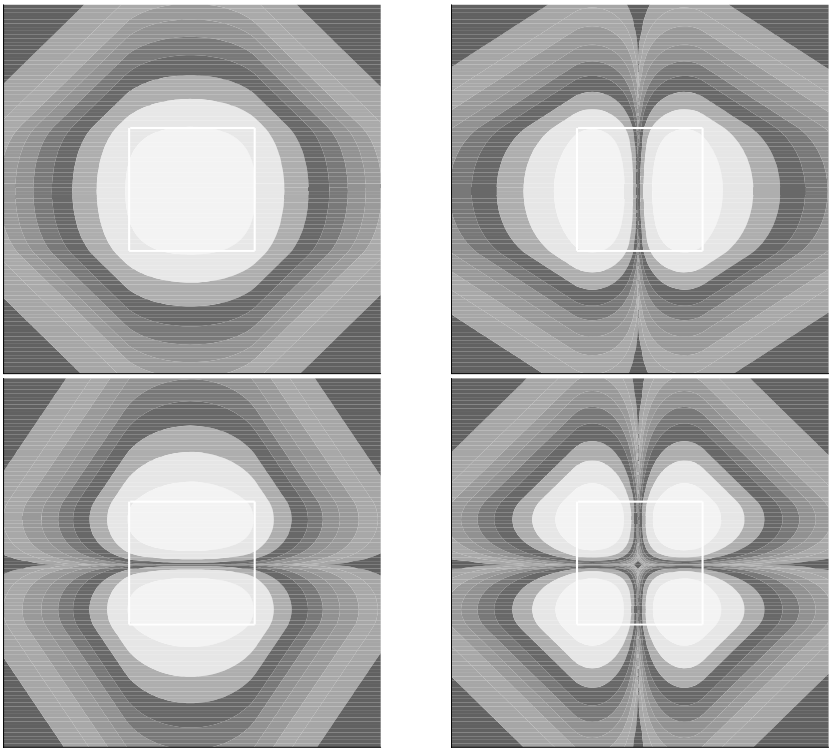


Figure 8.10: The charge densities of the four lowest-energy confined states of a GaAs quantum wire of side $100\,\text{Å}$ with finite $Ga_{0.8}Al_{0.2}As$ barriers, (top left) $n_y = 1$, $n_z = 1$; (top right) $n_y = 1$, $n_z = 2$; (bottom left) $n_y = 2$, $n_z = 1$; (bottom right) $n_y = 2$, $n_z = 2$; the edges of the wire are indicated by the boxes

theory, the change in energy of a level would be given by:

$$\Delta E = \langle \psi(y,z) | V'(y,z) | \psi(y,z) \rangle \tag{8.32}$$

i.e.

$$\Delta E = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \psi^*(y)\psi^*(z)V'(y,z)\psi(y)\psi(z)\,\mathrm{d}y\,\mathrm{d}z \tag{8.33}$$

The perturbation to the potential $V'(y,z)$ would be negative and of magnitude V, and employing the fourfold symmetry of this *square* cross-sectional wire would give:

$$\Delta E = -4V \int_{L_y}^{+\infty} \psi^*(y)\psi(y)\,\mathrm{d}y \int_{L_z}^{+\infty} \psi^*(z)\psi(z)\,\mathrm{d}z \tag{8.34}$$

which is relatively straightforward to evaluate. Califano and Harrison [7] have demonstrated that this can be a quite useful approach to the solution of finite-barrier quantum wires and dots.

8.5 Circular cross-section wire

Consider again the Schrödinger equation for the motion in the confined cross-sectional plane of a quantum wire, as given earlier in equation (8.11), i.e.

$$-\frac{\hbar^2}{2m^*}\left[\frac{\partial^2\psi(y,z)}{\partial y^2}+\frac{\partial^2\psi(y,z)}{\partial z^2}\right]+V(y,z)\psi(y,z)=E_{y,z}\psi(y,z) \tag{8.35}$$

Given the cylindrical symmetry of the quantum wire, as shown in Fig. 8.11, it would

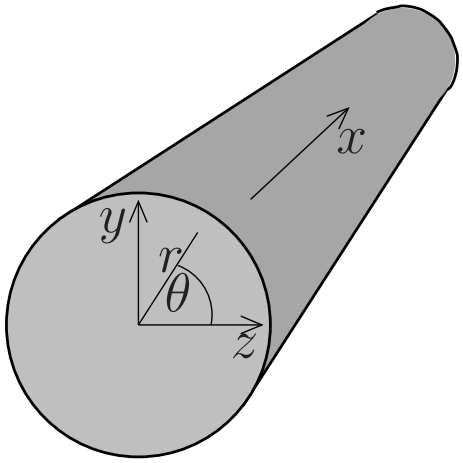


Figure 8.11: Schematic illustration of a circular cross-section quantum wire

seem advantageous to move into polar coordinates for the cross-sectional motion. With the definitions of the modulus r and angle θ as in the figure, the Cartesian coordinates then follow as:

$$y = r\sin\theta \quad \text{and} \quad z = r\cos\theta \tag{8.36}$$

$$\therefore r = \sqrt{y^2+z^2} \tag{8.37}$$

The wave function $\psi(y,z)$ can clearly be written in terms of the new variables r and θ.

The wave function, in general, could oscillate periodically over θ around the circular cross-section (see exercise (3)). However, the simplest solutions have wave functions that do not vary with θ, i.e. circularly symmetric (or s-like) functions. In this case, $\partial\psi/\partial\theta = 0$ and the wave function can be written as $\psi(r)$. The Schrödinger equation therefore becomes:

$$-\frac{\hbar^2}{2m^*}\left(\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\psi(r)+V(r)\psi(r)=E_r\psi(r) \tag{8.38}$$

where the index on E_r just indicates that this eigenvalue is associated with the confined cross-sectional motion, as opposed to the unconfined motion along the axis of the wire. In addition,

the circular symmetry of the potential that defines the wire can be written explicitly as $V(r)$. Now:

$$\frac{\partial}{\partial y}\psi(r) = \frac{\partial}{\partial r}\psi(r) \times \frac{\partial r}{\partial y} \tag{8.39}$$

Differentiating both sides of equation (8.37) with respect to y, gives:

$$\frac{\partial r}{\partial y} = \frac{1}{2}\left(y^2 + z^2\right)^{-\frac{1}{2}} \times 2y = \frac{y}{r} \tag{8.40}$$

Hence:

$$\frac{\partial}{\partial y}\psi(r) = \frac{\partial}{\partial r}\psi(r) \times \frac{y}{r} \tag{8.41}$$

The second derivative is then:

$$\frac{\partial}{\partial y}\frac{\partial}{\partial y}\psi(r) = \frac{\partial}{\partial y}\left[\frac{\partial}{\partial r}\psi(r) \times \frac{y}{r}\right] \tag{8.42}$$

$$\therefore \frac{\partial^2}{\partial y^2}\psi(r) = \frac{\partial^2}{\partial r^2}\psi(r) \times \frac{\partial r}{\partial y}\frac{y}{r} + \frac{\partial}{\partial r}\psi(r)\frac{\partial}{\partial y}\left(\frac{y}{r}\right) \tag{8.43}$$

and thus:

$$\frac{\partial^2}{\partial y^2}\psi(r) = \frac{y^2}{r^2}\frac{\partial^2}{\partial r^2}\psi(r) + \frac{\partial}{\partial r}\psi(r)\left(\frac{1}{r} - \frac{y}{r^2}\frac{\partial r}{\partial y}\right) \tag{8.44}$$

Finally:

$$\frac{\partial^2}{\partial y^2}\psi(r) = \frac{1}{r}\frac{\partial}{\partial r}\psi(r) - \frac{y^2}{r^3}\frac{\partial}{\partial r}\psi(r) + \frac{y^2}{r^2}\frac{\partial^2}{\partial r^2}\psi(r) \tag{8.45}$$

and similarly for z, hence:

$$\left(\frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\psi(r) = \frac{2}{r}\frac{\partial}{\partial r}\psi(r) - \frac{(y^2+z^2)}{r^3}\frac{\partial}{\partial r}\psi(r) + \frac{(y^2+z^2)}{r^2}\frac{\partial^2}{\partial r^2}\psi(r) \tag{8.46}$$

Recalling that $y^2 + z^2 = r^2$, then:

$$\left(\frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\psi(r) = \frac{1}{r}\frac{\partial}{\partial r}\psi(r) + \frac{\partial^2}{\partial r^2}\psi(r) \tag{8.47}$$

Substituting into equation (8.38) gives the final form for the Schrödinger equation as follows:

$$-\frac{\hbar^2}{2m^*}\left(\frac{1}{r}\frac{\partial}{\partial r} + \frac{\partial^2}{\partial r^2}\right)\psi(r) + V(r)\psi(r) = E_r\psi(r) \tag{8.48}$$

In this case, the specific form of the kinetic energy operator has been relied upon, unlike the earlier example of the rectangular cross-section quantum wire, and hence this Schrödinger equation *is* only valid for a constant effective mass.

One numerical approach for solving equation (8.48) would be to follow a similar procedure to that employed in Section 3.5, i.e. expand the derivatives in terms of the standard finite-difference equivalents:

$$\frac{\partial}{\partial r}\psi(r) = \frac{\psi(r+\delta r) - \psi(r-\delta r)}{2\delta r} \tag{8.49}$$

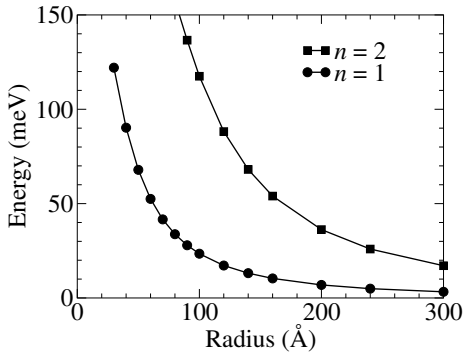


Figure 8.12: The confinement energy in a finite-barrier circular cross-section quantum wire

$$\frac{\partial^2}{\partial r^2}\psi(r) = \frac{\psi(r+\delta r) - 2\psi(r) + \psi(r-\delta r)}{(\delta r)^2} \tag{8.50}$$

and then rearrange to obtain a shooting iterative equation. This is achieved by substituting the above into equation (8.48) to give:

$$\frac{1}{r}\left[\frac{\psi(r+\delta r) - \psi(r-\delta r)}{2\delta r}\right] + \frac{\psi(r+\delta r) - 2\psi(r) + \psi(r-\delta r)}{(\delta r)^2} = \frac{2m^*}{\hbar^2}\left[V(r) - E_r\right]\psi(r) \tag{8.51}$$

Multiplying both sides by $2r(\delta r)^2$ gives:

$$\left[\psi(r+\delta r) - \psi(r-\delta r)\right]\delta r + 2r\left[\psi(r+\delta r) - 2\psi(r) + \psi(r-\delta r)\right]$$
$$= 2r(\delta r)^2\frac{2m^*}{\hbar^2}\left[V(r) - E_r\right]\psi(r) \tag{8.52}$$

Gathering terms in $\psi(r+\delta r)$, $\psi(r)$, and $\psi(r-\delta r)$:

$$(2r+\delta r)\,\psi(r+\delta r)$$
$$= 2r\left\{(\delta r)^2\frac{2m^*}{\hbar^2}\left[V(r) - E_r\right] + 2\right\}\psi(r) + (-2r+\delta r)\,\psi(r-\delta r) \tag{8.53}$$

Finally:

$$\psi(r+\delta r) = \frac{2r\left\{2m^*(\delta r/\hbar)^2\left[V(r) - E_r\right] + 2\right\}\psi(r) + (-2r+\delta r)\,\psi(r-\delta r)}{2r+\delta r} \tag{8.54}$$

which is an iterative shooting equation, similar to those met earlier in Chapter 3. It can be solved according to the standard boundary condition, $\psi(r) \to 0$ as $r \to \infty$, as before. Now in regions of constant potential, the wave functions are, in general, continuous, and therefore for the particular case of a straight line perpendicular to and through the wire (x-)axis, the wave function must also be continuous. Thus, when crossing the wire axis, the radial component

of the wave function $\psi(r)$ must have a derivative of zero, i.e. a local maximum or minimum. This allows the iterative starting conditions to be chosen as:

$$\psi(0) = 1 \quad \text{and} \quad \psi(\delta r) = 1 \tag{8.55}$$

These can look a little simplistic, but in the limit of decreasing step length δr, the maximum or minimum is clearly flat.

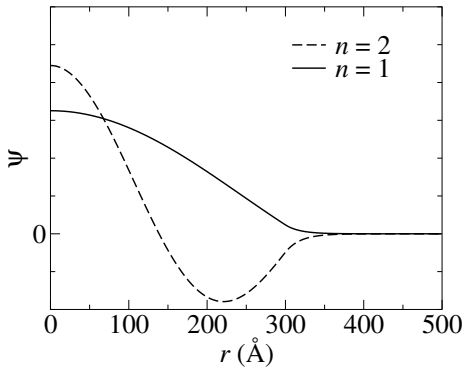


Figure 8.13: The radial component of the wave function $\psi(r)$ for the lowest two eigenstates in a finite-barrier quantum wire with radius 300 Å of circular cross-section

Figure 8.12 displays the results of calculations of the electron confinement energy versus the wire radius, for a GaAs wire surrounded by $Ga_{0.8}Al_{0.2}As$, *for constant effective mass*. As expected, the confinement energy decreases with increasing radius and the odd-parity eigenstate is of higher energy than the even. The latter point is highlighted in Fig. 8.13, which plots the radial motion $\psi(r)$ for the 300 Å radius wire. The even ($n = 1$) and odd ($n = 2$) parity nature of the eigenstates can clearly be seen.

8.6 Quantum boxes

Cuboid quantum dots, perhaps more specifically designated as *quantum boxes*, can be thought of as simply a generalisation of the rectangular cross-section quantum wires, in which there is now additional confinement along the remaining x-axis. This additional confinement removes the remaining degree of freedom in the particle's momentum and localises it in all directions. Thus the energy levels can no longer be referred to as *subbands* and are now known as *sublevels*.

Considering the case of an infinite potential separating the inside of the box from the outside, the three-dimensional Schrödinger equation within the box is simply:

$$-\frac{\hbar^2}{2m^*}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\psi(x,y,z) = E_{x,y,z}\psi(x,y,z) \tag{8.56}$$

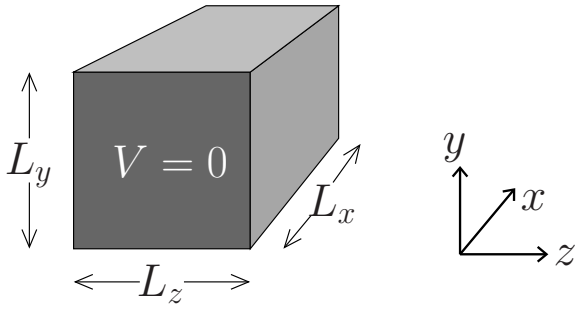


Figure 8.14: Schematic illustration of a quantum box with side L_x, L_y, and L_z

Again, writing the total energy $E_{x,y,z}$ as a sum of the three terms E_x, E_y, and E_z, this single three-dimensional equation can be decoupled into three one-dimensional equations:

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial x^2}\psi(x) = E_x\psi(x) \tag{8.57}$$

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial y^2}\psi(y) = E_y\psi(y) \tag{8.58}$$

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial z^2}\psi(z) = E_z\psi(z) \tag{8.59}$$

which, for the infinitely deep barrier case, are just the solutions to infinitely deep quantum wells, of widths L_x, L_y, and L_z respectively, i.e. in analogy to the infinitely deep quantum wire confinement energy of equation (8.25), the confinement energy within this quantum box follows as:

$$E_{x,y,z} = \frac{\hbar^2\pi^2}{2m^*}\left(\frac{n_x^2}{L_x^2} + \frac{n_y^2}{L_y^2} + \frac{n_z^2}{L_z^2}\right) \tag{8.60}$$

The three-dimensional nature of the confinement thus requires three *quantum numbers*, n_x, n_y and n_z, to label each state.

Finite-barrier quantum boxes could be gleaned from three decoupled one-dimensional quantum well calculations in a similar manner to the finite-barrier quantum wire discussed in Section 8.4. However, in this case any perturbative correction would have to account for eight corner cubes of additional potential $2V$ and 12 edge cuboids of additional potential V. Alternatively, a full three-dimensional solution can be constructed by expanding the wave function as a linear combination of infinite well solutions (see Gangopadhyay and Nag [8]).

8.7 Spherical quantum dots

It is perhaps easier to deal with a finite-barrier quantum dot with spherical rather than cuboid symmetry. The approach is rather similar to that derived earlier for the circular cross-section quantum wire.

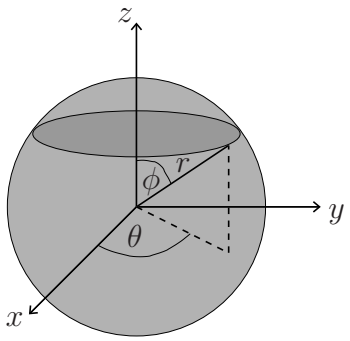


Figure 8.15: Schematic illustration of a spherical quantum dot

Given the spherical symmetry of the potential, the wave function (as was the case for a cylindrical wire) *can* have a periodic angular variation around the sphere, and these general solutions will be considered in Section 8.8. However, as before, the simplest (s-like) solutions do not vary with respect to angle. For these zero angular-momentum states, the Schrödinger equation for a *constant* effective mass could be written:

$$-\frac{\hbar^2}{2m^*}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\psi(r)+V(r)\psi(r)=E_r\psi(r) \tag{8.61}$$

where the index on E_r has been added just to indicate that this energy is associated with the confinement along the radius. In this case:

$$r=\sqrt{x^2+y^2+z^2} \tag{8.62}$$

The transition can be made from Cartesian (x, y, z) to spherical polar coordinates, in effect just r, in the same way as detailed in Section 8.5. Using equation (8.45), each of the three Cartesian axes gives an equation of the following form:

$$\frac{\partial^2}{\partial x^2}\psi(r)=\frac{1}{r}\frac{\partial}{\partial r}\psi(r)-\frac{x^2}{r^3}\frac{\partial}{\partial r}\psi(r)+\frac{x^2}{r^2}\frac{\partial^2}{\partial r^2}\psi(r) \tag{8.63}$$

Therefore, the complete $\nabla^2\psi(r)$ is given by:

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\psi(r)$$

$$=\frac{3}{r}\frac{\partial}{\partial r}\psi(r)-\frac{(x^2+y^2+z^2)}{r^3}\frac{\partial}{\partial r}\psi(r)+\frac{(x^2+y^2+z^2)}{r^2}\frac{\partial^2}{\partial r^2}\psi(r) \tag{8.64}$$

$$\therefore\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\psi(r)=\frac{2}{r}\frac{\partial}{\partial r}\psi(r)+\frac{\partial^2}{\partial r^2}\psi(r) \tag{8.65}$$

(as, for example, in [9], p. 188).

Substituting into the Schrödinger equation then:

$$-\frac{\hbar^2}{2m^*}\left(\frac{2}{r}\frac{\partial}{\partial r}+\frac{\partial^2}{\partial r^2}\right)\psi(r)+V(r)\psi(r)=E_r\psi(r) \tag{8.66}$$

Such spherically symmetric Schrödinger equations have been investigated before (see, for example, [10], p. 76). As an alternative to such a well-established approach, and with the impetus in this work on simple numerical schemes, a shooting technique similar to that described for the circular cross-section quantum wire is sought. With this aim, expanding the first and second derivatives in terms of finite differences gives:

$$\frac{2}{r}\left[\frac{\psi(r+\delta r)-\psi(r-\delta r)}{2\delta r}\right]+\frac{\psi(r+\delta r)-2\psi(r)+\psi(r-\delta r)}{(\delta r)^2} = \frac{2m^*}{\hbar^2}\left[V(r)-E_r\right]\psi(r) \tag{8.67}$$

Multiplying both sides by $r(\delta r)^2$ gives:

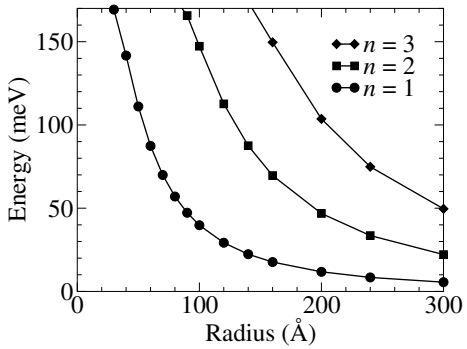


Figure 8.16: The confinement energy in a spherical GaAs quantum dot surrounded by a $Ga_{0.8}Al_{0.2}As$ barrier

$$\left[\psi(r+\delta r)-\psi(r-\delta r)\right]\delta r+r\left[\psi(r+\delta r)-2\psi(r)+\psi(r-\delta r)\right] = r(\delta r)^2\frac{2m^*}{\hbar^2}\left[V(r)-E_r\right]\psi(r) \tag{8.68}$$

Gathering terms in $\psi(r+\delta r)$, $\psi(r)$, and $\psi(r-\delta r)$:

$$(r+\delta r)\,\psi(r+\delta r) = r\left\{(\delta r)^2\frac{2m^*}{\hbar^2}\left[V(r)-E_r\right]+2\right\}\psi(r)+(-r+\delta r)\,\psi(r-\delta r) \tag{8.69}$$

and finally:

$$\psi(r+\delta r) = \frac{r\left\{2m^*(\delta r/\hbar)^2\left[V(r)-E_r\right]+2\right\}\psi(r)+(-r+\delta r)\,\psi(r-\delta r)}{r+\delta r} \tag{8.70}$$

which is again an iterative equation which can be solved with a numerical shooting technique according to the same boundary conditions as discussed in Section 8.5.

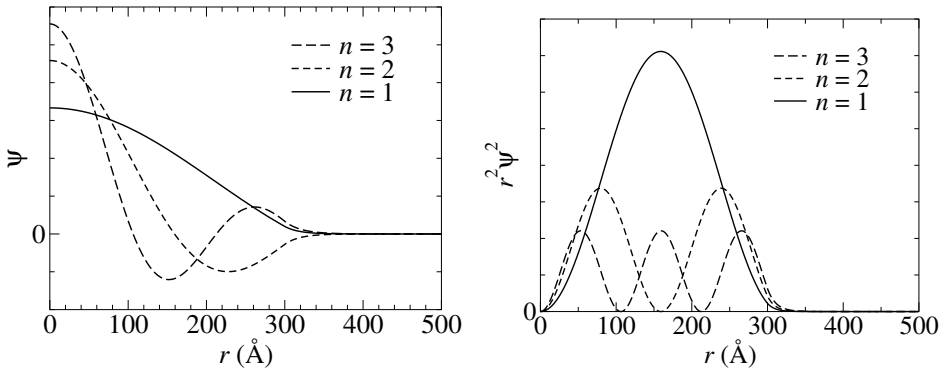


Figure 8.17: (a) The wave functions and (b) the probability densities of the three lowest-energy states in the 300 Å spherical quantum dot

Figure 8.16 shows the results of calculations of the three lowest energy levels of a spherical GaAs quantum dot surrounded by a finite barrier composed of $Ga_{0.8}Al_{0.2}As$, with a *sharp* boundary. In fact, the formalism above, as that of the circular cross-section quantum wire, is applicable for any radial potential profile $V(r)$, e.g. it is also valid for diffused interfaces. Again, the behaviour of the energies as a function of the spatial dimension, as shown in Fig. 8.16, is as expected in confined systems, namely the confinement energy decreases as the size of the system increases.

Figure 8.17(a) displays the corresponding radial components of the wave functions. It can be seen that they all have a maximum at the centre of the potential and that as the principal quantum number n increases, the number of nodes increases. The nature of the states is perhaps better illustrated by considering the charge density. Given that successive increases in the radial coordinate r lead to spherical shells of volume proportional to r^2, the charge density between the radii r and $r+\delta r$ is proportional to $r^2\psi^*(r)\psi(r)$; this is plotted in Fig. 8.17(b). The lowest-energy state, $n=1$, has a single antinode which is close to half the radius of the potential, while the second has two maxima in the charge density, etc.

8.8 Non-zero angular momentum states

For simplicity, the previous sections have ignored angular momentum, so in effect they have derived the radial components of the wave function for the zero angular momentum states.

Return now to 'del-squared' ∇^2 in spherical polar (r, θ, ϕ) coordinates:

$$\nabla^2 = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\phi^2} \tag{8.71}$$

Then limiting the case to spherically symmetric potentials $V(r)$, the Schrödinger equation would follow as:

$$-\frac{\hbar^2}{2m^*}\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right]\psi + V(r)\psi = E\psi \tag{8.72}$$

where it is still assumed the wave function ψ is a function of r, θ and ϕ. The standard approach, covered in any comprehensive book on quantum mechanics (e.g. [11]), is to attempt to separate the variables by writing:

$$\psi = \psi(r)Y(\theta,\phi) \tag{8.73}$$

Then equation (8.72) becomes:

$$-\frac{\hbar^2}{2m^*}\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi(r)}{\partial r}\right)\right]Y(\theta,\phi) + (V(r)-E)\,\psi(r)Y(\theta,\phi)$$

$$-\frac{\hbar^2}{2m^*}\left[\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial Y(\theta,\phi)}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 Y(\theta,\phi)}{\partial\phi^2}\right]\psi(r) = 0 \tag{8.74}$$

Multiplying by r^2 and dividing through by $\psi(r)Y(\theta,\phi)$ gives:

$$-\frac{\hbar^2}{2m^*}\left[\frac{1}{\psi(r)}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi(r)}{\partial r}\right)\right] + r^2\,(V(r)-E)$$

$$= \frac{\hbar^2}{2m^*}\frac{1}{Y(\theta,\phi)}\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial Y(\theta,\phi)}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2 Y(\theta,\phi)}{\partial\phi^2}\right] \tag{8.75}$$

Then:

$$\frac{1}{\psi(r)}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi(r)}{\partial r}\right) + \frac{2m^*r^2}{\hbar^2}\,(E-V(r))$$

$$= -\frac{1}{Y(\theta,\phi)}\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial Y(\theta,\phi)}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2 Y(\theta,\phi)}{\partial\phi^2}\right] \tag{8.76}$$

Now the left-hand side of the above equation is a function of r only and the right-hand side is a function of the angles θ and ϕ only, this can only be satisfied if both sides are equal to the same constant, call it λ, i.e.

$$\frac{1}{\psi(r)}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi(r)}{\partial r}\right) + \frac{2m^*r^2}{\hbar^2}\,(E-V(r)) = \lambda \tag{8.77}$$

and

$$-\frac{1}{Y(\theta,\phi)}\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial Y(\theta,\phi)}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2 Y(\theta,\phi)}{\partial\phi^2}\right] = \lambda \tag{8.78}$$

and hence the separation of variables is complete.

Following [10], the angular equation:

$$\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial Y(\theta,\phi)}{\partial\theta}\right)+\frac{1}{\sin^2\theta}\frac{\partial^2 Y(\theta,\phi)}{\partial\phi^2}+\lambda Y(\theta,\phi)=0 \tag{8.79}$$

can be further separated by writing $Y(\theta,\phi)=\Theta(\theta)\Phi(\phi)$. Identifying the constant of separation as ν leads to:

$$\frac{\mathrm{d}^2\Phi}{\mathrm{d}\phi}+\nu\Phi=0 \tag{8.80}$$

and

$$\frac{1}{\sin\theta}\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\sin\theta\frac{\mathrm{d}\Theta}{\mathrm{d}\theta}\right)+\left(\lambda-\frac{\nu}{\sin^2\theta}\right)\Theta=0 \tag{8.81}$$

Recalling standard boundary conditions, i.e. that the wave function is continuous, implies that $\Phi(\phi)$ is continuous at the boundary between 0 and 2π, hence ν needs to be positive—make it the square of an integer (m^2). Then the solution is:

$$\Phi(\phi)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{im\phi} \tag{8.82}$$

where different values of the integer m represent different solutions. Continuing without specifying the form of the potential $V(r)$ other than that it has spherical symmetry, progress continues along the standard lines for the solution of the hydrogen atom: the substitution $w=\cos\theta$ is made, together with writing $\Theta(\theta)=P(w)$. Then equation (8.81) becomes:

$$\frac{\mathrm{d}}{\mathrm{d}w}\left((1-w^2)\frac{\mathrm{d}P}{\mathrm{d}w}\right)+\left(\lambda-\frac{m^2}{1-w^2}\right)P=0 \tag{8.83}$$

Schiff [10] states that for $m=0$ the solutions are Legendre polynomials; for $m\neq 0$ acceptable solutions for $P(w)$ exist if $\lambda=l(l+1)$ and $|m|\leq l$. These solutions are known as associated Legendre functions.

The resulting angular solutions $Y(\theta,\phi)=\Theta(\theta)\Phi(\phi)$ are therefore written as:

$$Y_{lm}(\theta,\phi)=N_{lm}P_{lm}\Phi_m(\phi) \tag{8.84}$$

where l and m are the quantum numbers specifying the state and N_{lm} is a normalisation factor. These angular solutions are referred to as 'spherical harmonics'.

Returning to the radial equation (8.77), and recalling $\lambda=l(l+1)$, then:

$$\frac{1}{\psi(r)}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi(r)}{\partial r}\right)+\frac{2m^*r^2}{\hbar^2}\left(E-V(r)-\frac{l(l+1)\hbar^2}{2m^*r^2}\right)=0 \tag{8.85}$$

$$\therefore\left(2r\frac{\partial\psi(r)}{\partial r}+r^2\frac{\partial^2\psi(r)}{\partial r^2}\right)+\frac{2m^*r^2}{\hbar^2}\left(E-V(r)-\frac{l(l+1)\hbar^2}{2m^*r^2}\right)\psi(r)=0 \tag{8.86}$$

Divide by r^2 and multiply through by $-\hbar^2/2m^*$:

$$-\frac{\hbar^2}{2m^*}\left(\frac{2}{r}\frac{\partial\psi(r)}{\partial r}+\frac{\partial^2\psi(r)}{\partial r^2}\right)-\left(E-V(r)-\frac{l(l+1)\hbar^2}{2m^*r^2}\right)\psi(r)=0 \tag{8.87}$$

$$-\frac{\hbar^2}{2m^*}\left(\frac{2}{r}\frac{\partial\psi(r)}{\partial r}+\frac{\partial^2\psi(r)}{\partial r^2}\right)+\left(V(r)+\frac{l(l+1)\hbar^2}{2m^*r^2}\right)\psi(r)=E\psi(r) \tag{8.88}$$

which, for $l=0$, is the same as the earlier non-angular momentum radial equation (8.66). The term $l(l+1)\hbar^2/(2m^*r^2)$ looks like an additional energy associated with the angular momentum. The solution of this equation now requires the form of the potential $V(r)$ to be specified. One form would be to take $V(r)=-e^2/(4\pi\epsilon_0 r)$ which would, of course, lead to the solutions of the hydrogen atom. In the context here, a more appropriate choice could be for $V(r)=0$ for $0<r<a$ and for $V(r)=\infty$ or V_0 otherwise, which would define a spherical quantum dot of infinite or finite potential, respectively. This will be developed further in the exercises at the end of this chapter.

8.9 Approaches to pyramidal dots

There has been a drive to simplify the fabrication of quantum dots, and in a particular kind of material system this has been achieved with the aid of a process known as *self-assembly*. This occurs when a thin (perhaps even sub-mono) layer of one material is deposited on top of a substrate which has a quite different lattice constant. In such a system, the strain energy in the bonds forms a substantial fraction of the total energy, and in an effort to minimise this the deposited atoms rearrange themselves, moving from a thin layer into 'clumps'. This method has become known as the 'Stranski–Krastanov' growth mode [12]. Microscopy has shown that these small groups of atoms form pyramidal dots, as illustrated earlier in Fig. 8.4.

It appears that the pyramidal dot requires a fully three-dimensional solution of Schrödinger's equation, as given by Cusack *et al*. [13]. However, a little thought allows the problem to be simplified into a combination of a series of decoupled Schrödinger equations and some perturbation theory.

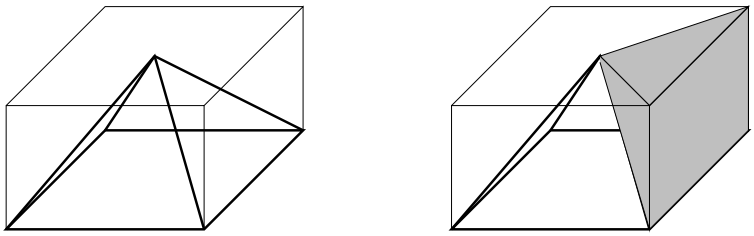


Figure 8.18: A possible perturbative approach for small pyramidal dots

For example, a pyramidal dot could be thought of as a finite-barrier cuboid dot, but with four additional perturbing potentials, as shown in Fig. 8.18 (right) (shaded area). These also take the shape of square based pyramids, with the base, in this example, vertical. However, these pyramids are not symmetric, having the apex centred on the upper edge rather than over the centre of the base.

If the finite dot was solved with decoupled one-dimensional wave functions, then for relatively small systems (less than 100 Å) where the unperturbed wave function penetrates

considerably into the barrier of the cuboid quantum dot, there is the possibility of a partial cancellation of the $-2V$ and $-V$ potentials at the corners and along the edges (see Section 8.6) with these $+V$ non-symmetric pyramidal potentials. Such a partial cancellation may be vital to ensure convergence of the perturbed energy levels. By employing this approach, connection rules have been derived between the energy of a pyramidal dot and a cubic quantum dot of a different volume [7].

As mentioned above, lattice mismatch is what drives the self-assembly process, hence the atoms constituting self-assembled quantum dots are under the influence of significant amounts of strain. The effect of this strain is to alter the conduction- and valence-band profiles so they are no longer constant either inside or outside the quantum dot. The method for determining the strain field and the effect on the conduction- and valence-band edges has been summarised by Califano and Harrison [14] for both $In_{1-x}Ga_xAs$ on GaAs and $Si_{1-x}Ge_x$ on silicon quantum dots.

8.10 Matrix approaches

Alternative solutions are being sought to this problem by using an extension of the method of Gangopadhyay and Nag cited earlier [8]. This involves expanding the wave function as a Fourier series of infinite quantum well solutions and has proved itself to be a workable and reliable technique [15, 16].

The basic approach is to expand the two- or three-dimensional wave function of the quantum wire or quantum dot as a linear combination of some basis functions. The basis functions are usually sine or plane waves of a larger box that encompasses the quantum wire or dot and has an infinite exterior potential. This leads to a formulation where the energy solutions are the eigenvalues of a matrix equation. There is a thorough introduction to this method and a description of how to apply it to the solution of Schrödinger's equation in two and three dimensions in Chapter 3 of the book by Harrison [17], and the next chapter is focused on applying it to quantum dots of arbitrary shape.

8.11 Finite-difference expansions

Although less important with the advent of desktop computers with more and more memory, the main problem with matrix methods for the solution of Schrödinger's equation in multiple dimensions is the sheer size of the expansion set. If in one dimension it was found that ten basis states were needed to reproduce the ground state energy and wave function of some perturbed potential, then in two dimensions for similar accuracy the basis set would have $10 \times 10 = 100$ components. In three dimensions this would become $10 \times 10 \times 10 = 1000$ states. Thus the matrix would be of order 100 or 1000, respectively and this is for a relatively small number of basis functions; in reality 20 or more in each direction may be necessary.

An alternative approach [18, 19] which is less demanding on memory is to return to the idea of a finite-difference expansion of the derivatives in Schrödinger's equation, similar to that developed for the shooting method for one dimension in Chapter 3. Recall the Schrödinger equation for the motion in the confined cross-sectional plane of a quantum wire,

as given earlier in equation (8.11) and used again in equation (8.35):

$$-\frac{\hbar^2}{2m^*}\left[\frac{\partial^2\psi(y,z)}{\partial y^2}+\frac{\partial^2\psi(y,z)}{\partial z^2}\right]+V(y,z)\psi(y,z)=E_{y,z}\psi(y,z) \tag{8.89}$$

Expanding the two derivatives in terms of finite differences, as in equation (3.15), gives:

$$\begin{aligned}-\frac{\hbar^2}{2m^*}&\left[\frac{\psi(y+\delta y,z)-2\psi(y,z)+\psi(y-\delta y,z)}{(\delta y)^2}\right.\\ &\left.+\frac{\psi(y,z+\delta z)-2\psi(y,z)+\psi(y,z-\delta z)}{(\delta z)^2}\right]+V(y,z)\psi(y,z)=E_{y,z}\psi(y,z)\end{aligned} \tag{8.90}$$

Multiplying through by $(\delta y)^2(\delta z)^2$:

$$\begin{aligned}&(\delta z)^2\left[\psi(y+\delta y,z)-2\psi(y,z)+\psi(y-\delta y,z)\right]\\ &+(\delta y)^2\left[\psi(y,z+\delta z)-2\psi(y,z)+\psi(y,z-\delta z)\right]\\ &=-\frac{2m^*}{\hbar^2}\left[E_{y,z}-V(y,z)\right](\delta y\delta z)^2\,\psi(y,z)\end{aligned} \tag{8.91}$$

Gathering terms together:

$$\begin{aligned}&(\delta z)^2\left[\psi(y+\delta y,z)+\psi(y-\delta y,z)\right]+(\delta y)^2\left[\psi(y,z+\delta z)+\psi(y,z-\delta z)\right]\\ &+\left\{\frac{2m^*}{\hbar^2}\left[E_{y,z}-V(y,z)\right](\delta y\delta z)^2-2\left[(\delta y)^2+(\delta z)^2\right]\right\}\psi(y,z)=0\end{aligned} \tag{8.92}$$

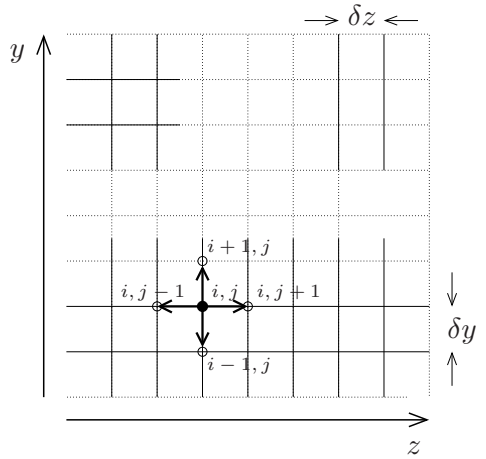


Figure 8.19: The two-dimensional mesh for the finite-difference approach to the quantum wire

Thus the wave function at the general point $\psi(y,z)$ depends upon the values at four neighbouring points (see Fig. 8.19). Equation (8.92) can be written more succinctly in terms of the indices given in Fig. 8.19:

$$(\delta y)^2(\psi_{i,j+1}+\psi_{i,j-1})+(\delta z)^2(\psi_{i+1,j}+\psi_{i-1,j})+k\psi_{i,j}=0 \tag{8.93}$$

where:

$$k=\frac{2m^*}{\hbar^2}\left[E_{y,z}-V(y,z)\right](\delta y\delta z)^2-2\left[(\delta y)^2+(\delta z)^2\right] \tag{8.94}$$

Such an equation exists for each grid point on the mesh, and to solve simultaneously they have to be written in the form of a matrix equation and solved according to the standard boundary conditions of the wave function and its first derivative tending to zero as the spatial coordinates tend to infinity. The memory saving occurs because the matrix is sparse. El-Moghraby *et al*. [18] map out this method in detail and apply it to rectangular and triangular cross-section quantum wires and pyramidal self-assembled quantum dots. In a later work El-Moghraby *et al*. [19] apply the method to vertically aligned coupled quantum dots.

8.12 Density of states

Just as there is a change in the density of states moving from the bulk (3D) crystal to a quantum well (2D), there is a further change in the density of states on moving to quantum wires (1D) and quantum dots (0D). Recall from Section 2.4 that the density of states is defined as the number of states per unit energy per unit volume of real space, which was expressed mathematically in equation (2.42) as:

$$\rho(E)=\frac{\mathrm{d}N}{\mathrm{d}E} \tag{8.95}$$

In the bulk crystal, the three degrees of freedom for the electron momentum mapped out a sphere in **k**-space, while in a quantum well the electron momenta fill successively larger circles. Continuing this argument for a quantum wire with just one degree of freedom, the electron momenta then fill states along a line.

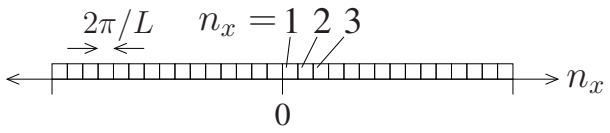


Figure 8.20: The occupation of states in **k**-space along a one-dimensional quantum wire

Therefore, proceeding with the same argument as before, the total number of states N is equal to the length of the line in **k**-space ($2k$), divided by the length occupied by one state (again $2\pi/L$), and divided by the *length* in real space (see Fig. 8.20), i.e.

$$N^{1\mathrm{D}}=2\times 2k\,\frac{1}{2\pi/L}\,\frac{1}{L} \tag{8.96}$$

where again the prefactor 2 accounts for spin degeneracy. Therefore:

$$N^{1D} = \frac{4k}{2\pi} \tag{8.97}$$

and for later:

$$\frac{dN^{1D}}{dk} = \frac{2}{\pi} \tag{8.98}$$

In analogy to both the bulk and quantum well cases, define the density of states for a one-dimensional wire as:

$$\rho^{1D}(E) = \frac{dN^{1D}}{dE} = \frac{dN^{1D}}{dk}\,\frac{dk}{dE} \tag{8.99}$$

As the along-axis dispersion curve can still be described as a parabola (as given by equation (8.12)), equation (2.48) can be reused, i.e.

$$\frac{dk}{dE} = \left(\frac{2m^*}{\hbar^2}\right)^{\frac{1}{2}} \frac{E^{-\frac{1}{2}}}{2} \tag{8.100}$$

This, taken together with the result in equation (8.98), then gives:

$$\rho^{1D}(E) = \frac{2}{\pi}\,\left(\frac{2m^*}{\hbar^2}\right)^{\frac{1}{2}} \frac{E^{-\frac{1}{2}}}{2} \tag{8.101}$$

$$\therefore \rho^{1D}(E) = \left(\frac{2m^*}{\hbar^2}\right)^{\frac{1}{2}} \frac{1}{\pi E^{\frac{1}{2}}} \tag{8.102}$$

where the energy E is measured upwards from a subband minimum. Therefore, comparing the density of states for bulk (3D), quantum wells (2D) and quantum wires (1D), as summarised in Table 8.2, it can be seen that successive reductions in degrees of freedom for the electron motion lead to reductions in the functional form of $\rho(E)$ by factors of $E^{\frac{1}{2}}$.

Table 8.2 The density of states for reduced-dimensionality systems, rewritten in a standard form

Dimensionality	$\rho(E)$	
3D	$\frac{1}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{\frac{3}{2}}$	$E^{\frac{1}{2}}$
2D	$\frac{1}{2\pi}\left(\frac{2m^*}{\hbar^2}\right)^{1}$	E^{0}
1D	$\frac{1}{\pi}\left(\frac{2m^*}{\hbar^2}\right)^{\frac{1}{2}}$	$E^{-\frac{1}{2}}$

If there are many (n) confined states within the quantum wire with subband minima E_i, then the density of states at any particular energy is the sum over all the subbands below that point, which can be written as:

$$\rho^{1D}(E) = \sum_{i=1}^{n} \left(\frac{2m^*}{\hbar^2}\right)^{\frac{1}{2}} \frac{1}{\pi(E-E_i)^{\frac{1}{2}}}\Theta(E-E_i) \tag{8.103}$$

Figure 8.21 gives an example of the 1D density of states, for a 60×70 Å rectangular cross-section GaAs quantum wire surrounded by infinite barriers. In contrast to the bulk and 2D cases displayed in Fig. 2.9, quantum wires show maxima in the density of states at around the subband minima, i.e. at around the point at which charge would be expected to accumulate. Therefore, interband (electron–hole) recombination will have a narrower linewidth than that of the 2D or 3D case.

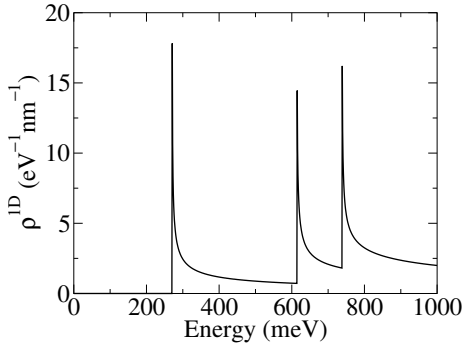


Figure 8.21: The density of states in a one-dimensional quantum wire

The situation for quantum dots is quite different. As the particles are confined in all directions, then there are no dispersion curves, and thus the density of states is just dependent upon the number of confined levels. One single isolated dot would therefore offer just two (spin-degenerate) states at the energy of each confined level, and a plot of the density of states versus energy would be a series of δ-functions.

Exercises

(1). By considering the confinement energy in a rectangular cross-section quantum wire surrounded by infinitely high barriers, show that for any given cross-sectional area the minimum confinement energy occurs when the cross-section is square.

(2). Section 8.4 approximates the solution of a rectangular cross-section quantum wire by assuming the wave function can be separated in the two directions y and z perpendicular to the axis of the quantum wire. Assuming constant effective mass m^*, use first-order perturbation theory to deduce an expression for the error this method introduces for the ground state energy of an electron in a square cross-section quantum wire surrounded by barriers of height V. (Hint: putting the origin at the centre of the quantum wire allows simplifications through symmetry.)

(3). An electron in a circular cross-section quantum wire surrounded by an *infinitely* high confining potential has an analytical form for its wave function in the single-band effective mass model. By converting the three-dimensional Schrödinger equation into

cylindrical polar coordinates, taking a form for the wave function as:

$$\psi = \psi(r)\Theta(\theta)\mathrm{e}^{ikz}$$

and separating the variables, deduce the analytical form for the components $\psi(r)$ and $\Theta(\theta)$ of the wave function. Find an expression for the corresponding energy eigenvalue.

(4). Given that InAs forms a zinc blende type crystal structure with a lattice constant of 6.06 Å, calculate the number of *atoms* in a cuboidal quantum dot of size L_x, L_y, L_z.

(5). Deduce the analytical form for the *radial* component $\psi(r)$ of the wave functions of a spherical quantum dot of radius a with an infinitely high confining potential, and find an expression for the corresponding energy eigenvalue. (Hint: start with equation (8.88).)

(6). In Section 2.4, expressions were derived for the density of states in bulk materials and quantum wells with non-parabolic energy bands. Find an equivalent expression for a quantum wire, and determine the form of the energy-dependent density-of-states effective mass.

References

[1] M. J. Kelly, *Low Dimensional Semiconductors: Materials, Physics, Technology, Devices*, Clarendon Press, Oxford, 1995.

[2] S. Gangopadhyay and B. R. Nag, 'Energy levels in finite barrier triangular and arrowhead-shaped quantum wires', *J. Appl. Phys.*, **81**:7885, 1997.

[3] R. Turton, *The Quantum Dot: a Journey into the Future of Microelectronics*, Oxford University Press, New York, 1995.

[4] I. Kamiya, I. Tanaka, and H. Sakaki, 'Optical properties of near surface-InAs quantum dots and their formation processes', *Physica E*, **2**:637, 1998.

[5] E. Palange, G. Capellini, L. D. Gaspare, and F. Evangelisti, 'Atomic force microscopy and photoluminescence study of Ge layers and self-organised Ge quantum dots on Si(100)', *Appl. Phys. Lett.*, **68**:2982, 1996.

[6] D. Gershoni, H. Temkin, G. J. Dolan, J. Dunsmuir, S. N. G. Chu, and M. B. Panish, 'Effects of two-dimensional confinement on the optical properties of InGaAs/InP quantum wire structures', *Appl. Phys. Lett.*, **53**(11):995–997, 1988.

[7] M. Califano and P. Harrison, 'Approximate methods for the solution of quantum wires and dots: connection rules between pyramidal, cuboid and cubic dots', *J. Appl. Phys.*, **86**:5054–5059, 1999.

[8] S. Gangopadhyay and B. R. Nag, 'Energy levels in three-dimensional quantum-confinement structures', *Nanotechnology*, **8**(1):14, 1997.

[9] R. T. Weidner and R. L. Sells, *Elementary Modern Physics*, Allyn and Bacon, Boston, Third edition, 1980.

[10] L. I. Schiff, *Quantum Mechanics*, McGraw-Hill, New York, 1968.

[11] R. L. Liboff, *Introductory Quantum Mechanics*, Addison–Wesley, San Francisco, Fourth edition, 2003.

[12] D. Bimberg, M. Grundmann, and N. N. Ledentsov, *Quantum Dot Heterostructures*, John Wiley & Sons, Ltd., Chichester, 1999.

[13] M. A. Cusack, P. R. Briddon, and M. Jaros, 'Electronic structure of InAs/GaAs self-assembled quantum dots', *Phys. Rev. B*, **54**:R2300–R2303, 1996.

[14] M. Califano and P. Harrison, 'Composition, volume and aspect ratio dependence of the strain distribution, band lineups and electron effective masses in self-assembled pyramidal $In_{1-x}Ga_xAs/GaAs$ and Si_xGe_{1-x}/Si quantum dots', *J. Appl. Phys.*, **91**:389–398, 2002.

[15] M. Califano and P. Harrison, 'Presentation and experimental validation of a single-band, constant-potential model for self-assembled InAs/GaAs quantum dots', *Phys. Rev. B*, **61**(16):10959–10965, 2000.

[16] M. Califano and P. Harrison, 'Quantum box energies as a route to the ground state levels of self-assembled quantum dots', *J. Appl. Phys.*, **88**:5870–5874, 2000.

[17] P. Harrison, *Computational Methods in Physics, Chemistry and Mathematical Biology: An Introduction*, John Wiley & Sons, Ltd, Chichester, 2001.

[18] D. El-Moghraby, R. G. Johnson, and P. Harrison, 'Calculating modes of quantum wire and dot systems using a finite differencing technique', *Comp. Phys. Commun.*, **150**:235–246, 2003.

[19] D. El-Moghraby, R. G. Johnson, and P. Harrison, 'The effect of inter-dot separation on the finite difference solution of vertically aligned coupled quantum dots', *Comp. Phys. Commun.*, **155**:236–243, 2003.

9

Quantum dots

M. Califano

University of Leeds

9.1 Zero-dimensional systems and their experimental realisation

As discussed in Chapter 8, structures in which the the motion of the carriers is confined further by increasing the value of $\mathcal{D}_c$ to 3 (i.e. confinement in all three dimensions) are called *quantum dots*. Because the number of degrees of freedom $\mathcal{D}_f$ in these structures is 0 (see equation (8.1) and Table 8.1) they are also known as *zero-dimensional* systems.

A quantum dot can be thought of as an artificial atom, as its density of states consists of a series of very sharp peaks, and its physical properties resemble, in many respects, those of an atom in a cage. However, these artificial atoms are also expected to have some significantly different properties compared with real atoms, as they can be filled with both electrons and holes.

Historically, the first techniques employed to realise semiconductor heterostructures that provided 3D confinement for the carriers were lithographic patterning and etching of quantum well structures. There are, in practice, different lithographic techniques (optical lithography and holography, X-ray lithography, electron and focused ion beam lithography, scanning tunnelling lithography, etc.). The fundamental steps involved are, however, essentially the same: (i) the growth of a layered structure followed by the imposition of a further structure on that, or (ii) the formation of a three-dimensional pattern during growth.

These techniques involve serial processes and are therefore very slow. Nevertheless, they also have several advantages that still make them appealing, such as the possibility of realising dots of almost arbitrary lateral shape, size and arrangement, depending on the resolution of the particular lithographic technique used, and their general compatibility with modern very large-scale integrated semiconductor technology.

Another related method worth mentioning consists of the creation (with lithographic techniques) of miniature electrodes on the surface of a quantum well. A spatially modulated electric field, which localises the electrons within a small region, is then produced by

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

applying a suitable voltage to the electrodes. Such electrostatically confined dots are also called *parabolic dots*, due to the fact that the shape of the confining potential can be approximated reasonably well by a parabola [1].

A different approach to the creation of quantum dot structures is self-assembly. This can be achieved in two ways: by chemical or by epitaxial synthesis.

Chemical synthetic processes provide a cheap and fast technique by which almost perfectly crystalline clusters ranging from few hundreds to tens of thousands of atoms can be obtained in a variety of shapes (spheres, rods, arrows, tetrapods, trees and so on) and materials. These so-called *nanocrystals* or *colloidal dots* are synthesised in solution and are covered by organic molecules (ligands) that allow them to be soluble and prevent aggregation. They can be dried and deposited onto substrates made of other materials, incorporated into devices such as photovoltaic cells as sensitisers or used as markers in biological applications.

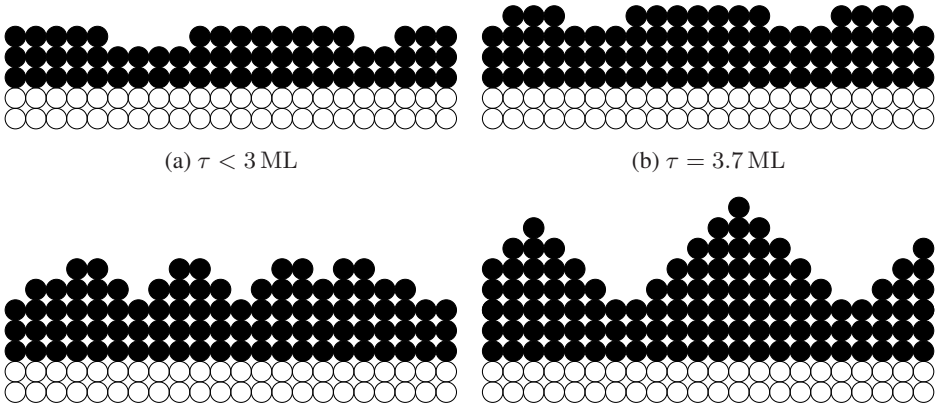
In the epitaxial deposition on each other of semiconductor materials that differ slightly in lattice constant, the resulting lattice strain is exploited to obtain arrays of three-dimensional islands or quantum dots. In fact, when a material is grown on top of another material with a different lattice constant, as is the case for InGaAs on GaAs, it can strain to conform to the substrate in the plane of the junction (it is assumed that the substrate is so thick that it cannot be distorted significantly). Thus its lattice constant in that plane is reduced (for InGaAs) while it is extended along the direction of growth (the opposite occurs for Si on SiGe) due to the elastic response of the material (as illustrated in Fig. 7.3). This distortion of the active layer causes the build-up of elastic energy, which is relaxed when a critical thickness τ, which depends on the particular heterostructure, is reached, and the two-dimensional growth changes into a three-dimensional one, as shown schematically in Fig. 9.1 for the case of SiGe: coherent [2, 3] (i.e. defect-free) islands of the deposited material form spontaneously, with a thin wetting layer left under the islands. The quantum dot islands are then covered (capped, in technical terms) with a layer of the substrate material.

Islands of various sizes and shapes have been reported, depending on the growth conditions: square-based pyramids with typical base length and height of 120 and $60\,\text{Å}$ [4], or 240 and $30\,\text{Å}$ [5] respectively, lens-shaped (hemispherical) islands with base diameters and heights of 200 and $50\,\text{Å}$ [6], or 200 and $7\,\text{Å}$ [7], respectively. Other islands have been observed [8] to have a square base pyramidal shape before GaAs regrowth, and a lens shape after that, or to become truncated pyramids with an In-rich core having an inverted-triangle [9] or an inverted-cone shape after capping.

It is worth stressing that both shape and size are important parameters, variations in which can dramatically affect the electronic structure of the island in both conduction and valence bands. Therefore, their knowledge is essential in order for a theoretical model to give an accurate description of the electronic and optical properties of a single quantum dot.

The complexity of the problem increases further, as it is now possible, by using suitable growth conditions, to fabricate arrays and stacks of small self-assembled quantum dots ($\sim 10\,\text{nm}$), ordered in size and shape, with high area density (greater than $10^{11}\,\text{cm}^{-2}$) and high optical quality, suitable for creating devices such as low threshold lasers [10]. Furthermore, the presence of external electric and/or magnetic fields is often required to control the number of charge carriers inside these structures.

Calculating the energy levels of a *single* quantum dot alone is already a formidable computational task and the Schrödinger equation must often be solved in three dimensions



(a) $\tau < 3\,\mathrm{ML}$ (b) $\tau = 3.7\,\mathrm{ML}$

(c) $\tau > 3.7\,\mathrm{ML}$ (d) $\tau \gg 3.7\,\mathrm{ML}$

Figure 9.1: Schematic illustration of the growth of Ge (solid circles) on a Si substrate (open circles). The initial (a) 2D layer-by-layer growth continues up to the critical thickness for the island formation ($\tau = 3.7$ monolayers (ML)) (b), after which small islands nucleate on the flat 3.7 ML Ge (c), which turn into 3D islands (d) leaving 3.0 ML of 2D Ge

by means of a numerical method, where in most cases recourse is made to approximations and simplifying hypotheses such as the assumption of equal and isotropic elastic properties of both dot and matrix, and of an effective mass variable only with composition but not with position [11] (the position dependence of the strain components leads, in fact, to a position-dependent effective mass [12]), invariance under strain of the matrix elements in the calculation of the electron effective masses [13], decoupling of valence and conduction bands in the calculation of the energies and wave functions [14], neglect of spin–orbit split-off band coupling effects [13, 14], etc. Considered in this light, the problem of solving the Schrödinger equation for an ensemble of dots in a complex structure becomes almost hopeless if all sorts of contributions have to be considered properly. The need therefore arises for as simple a model as possible for describing the electronic structure of a single dot.

9.2 Cuboidal dots

One of the easiest dot shapes to consider in Cartesian coordinates is the cube. A cuboidal dot can, in fact, be thought of as a generalisation of the rectangular cross-section quantum wire treated in Chapter 8, where the motion of the carriers is further confined along the x-axis. Quantum boxes were treated in Section 8.6; however, due to their importance for what follows, their solutions will be presented here as well.

In the simplest case of an infinite confining potential, within the dot the three-dimensional Schrödinger equation (8.6) assumes the form of equation (8.56). The eigenstates of this system will be exploited later on in this chapter as a basis set for the expansion of the solution of the more general problem of a dot with arbitrary shape in a finite potential. By writing the

total wave function as a product, and the energy as a sum of three terms:

$$\psi(x,y,z) = \psi(x)\psi(y)\psi(z) \tag{9.1}$$

$$E_{x,y,z} = E_x + E_y + E_z \tag{9.2}$$

equation (8.56) can be decoupled into a set of three independent one-dimensional equations, following the same procedure used in Section 8.3, yielding equations (8.57), (8.58) and (8.59), each of which is identical to equation (2.6), relative to an infinitely deep quantum well (here the three wells are orientated along the three perpendicular directions x, y and z). The total solutions of equation (8.3) will therefore be products of three terms of the form of equation (2.15) in all directions:

$$\psi_{lmn} = \sqrt{\frac{2}{L_x}}\sin\left(\frac{l\pi x}{L_x}\right)\sqrt{\frac{2}{L_y}}\sin\left(\frac{m\pi y}{L_y}\right)\sqrt{\frac{2}{L_z}}\sin\left(\frac{n\pi z}{L_z}\right), \tag{9.3}$$

and the energies of the confined states will be as in equation (8.60), i.e. sums of three terms like equation (2.13):

$$E_{x,y,z} = \frac{\hbar^2\pi^2}{2m^*}\left(\frac{n_x^2}{L_x^2} + \frac{n_y^2}{L_y^2} + \frac{n_z^2}{L_z^2}\right). \tag{9.4}$$

9.3 Dots of arbitrary shape

In the more interesting case of a realistic (i.e. finite) potential, $V(x,y,z)$ cannot be expressed as a sum of three independent potentials $V(x)$, $V(y)$ and $V(z)$, therefore the three-dimensional equation of motion cannot be decoupled into three independent one-dimensional equations. Also, as was mentioned at the beginning of this chapter, the shape of the dots synthesised experimentally may vary greatly from the idealised cube treated in the previous section. Still, it is possible to develop a simple model to describe their electronic and optical properties, within the framework of the effective mass approximation using a single-band envelope function technique.

Gershoni *et al*. [15] first developed a similar approach in 2D, in which they expressed the envelope function Ψ of a rectangular quantum wire with finite confining potential in terms of a linear combination of periodic functions ψ (called basis functions), solutions of a rectangular wire with an infinite barrier height and suitably chosen dimensions:

$$\Psi(x,y,z) = \sum_{lmn} a_{lmn}\psi_{lmn}(x,y,z). \tag{9.5}$$

This expansion is mathematically equivalent to the function Ψ, as the basis functions $\{\psi\}$ form a complete and orthonormal set (COS). However, the choice of the basis set is not unique, and many different COSs can be used. The selection of a specific basis is usually based on symmetry considerations, aiming to minimise the number of functions required to describe a specific system (the larger this number, the longer the calculation will take to run and the larger the memory requirement will be!). In fact, as any COS $\{\psi\}$ is infinite, for practical purposes (i.e. to be able to perform any calculation), only a limited

number of basis functions are used and the set is truncated at some combination of values of l, m, n corresponding to a cut-off wave vector $\mathbf{k}_{\text{cut}} = (k_x, k_y, k_z)$ in reciprocal space, see equation (2.12). The error made in this approximation needs, however, to be assessed by running convergence tests in which energy eigenvalues and other relevant quantities (e.g. dipole matrix elements) are calculated (see below). The accuracy can be increased by increasing the number of states included in the expansion (equation (9.5)).

The advantage of this technique is that, in contrast to the case of the finite quantum well presented in Section 2.7, here there is no need to explicitly match the wave functions across the boundary between barrier and dot materials, as the basis functions extend over the whole system. This method can therefore be applied to structures of arbitrary shape. Moreover, all the matrix elements can be calculated analytically, allowing considerable savings in computational resources.

Gangopadhyay and Nag [16] extended this method to study 3D confined structures such as parallelepipeds and cylinders, whereas Califano and Harrison [12] applied it to pyramidal structures.

In the effective mass approximation the Schrödinger equation for the envelope function of an arbitrarily shaped dot can be written as:

$$-\frac{\hbar^2}{2}\left(\nabla\frac{1}{m^*(x,y,z)}\nabla\right)\Psi(x,y,z) + V(x,y,z)\Psi(x,y,z) = E\Psi(x,y,z) \tag{9.6}$$

This form ensures, among other things, that (i) the Hamiltonian is Hermitian, (ii) the wave functions are orthogonal and (iii) the probability current is conserved at the interface of the heterojunction. The envelope function of the dot, $\Psi(x, y, z)$, can then be expanded in terms of a COS of solutions of a structure with infinite confining potential and suitable symmetry (sphere, cylinder, cube, etc.) to match that of the dot. The following example will use as a basis set the solutions of the *cuboidal* problem with infinite barrier height deduced above, i.e.

$$\psi_{lmn} = \sqrt{\frac{2}{L_x}}\sin\left[l\pi\left(\frac{1}{2}-\frac{x}{L_x}\right)\right]\sqrt{\frac{2}{L_y}}\sin\left[m\pi\left(\frac{1}{2}-\frac{y}{L_y}\right)\right]\sqrt{\frac{2}{L_z}}\sin\left[n\pi\frac{z}{L_z}\right] \tag{9.7}$$

where the domains $[-L_x/2, L_x/2]$, $[-L_y/2, L_y/2]$ were chosen for the variation of x and y, and $[0, L_z]$ for that of z (hence the slight difference in the argument of the first two sine functions from those in equation (9.3) above).

Care has to be taken to place the boundaries L_x, L_y, L_z away from the dot surface, so that the energy eigenvalues are independent of their choice. This can be achieved by solving equation (9.6) for different (increasing) sizes of the outer box until the energies obtained converge (see Section 9.3.1 below). As mentioned before, one of the advantages of this approach is that, because there is no need to explicitly match wave functions across the boundary between barrier and dot, it can easily be applied to an arbitrary confining potential. Substituting expression (9.5) into equation (9.6) leads to:

$$\begin{aligned}&-\frac{\hbar^2}{2}\nabla\left(\frac{1}{m^*(x,y,z)}\nabla\sum_{lmn}a_{lmn}\psi_{lmn}(x,y,z)\right)\\&+V(x,y,z)\sum_{lmn}a_{lmn}\psi_{lmn}(x,y,z) = E\sum_{lmn}a_{lmn}\psi_{lmn}(x,y,z)\end{aligned} \tag{9.8}$$

Multiplying on the left by $\psi^*_{l'm'n'}$, and integrating over the cuboid $L_xL_yL_z$ yields:

$$\sum_{lmn} a_{lmn} \left[-\frac{\hbar^2}{2} \int \psi^*_{l'm'n'} \nabla \left(\frac{1}{m^*(x,y,z)} \nabla \psi_{lmn} \right) \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z \right.$$

$$\left. + \int \psi^*_{l'm'n'} V \psi_{lmn} \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z - E \int \psi^*_{l'm'n'} \psi_{lmn} \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z \right] = 0 \tag{9.9}$$

which can be expressed in the form of a matrix equation as:

$$(M_{mlnm'l'n'} - E\delta_{mm'}\delta_{ll'}\delta_{nn'})a_{lmn} = 0 \tag{9.10}$$

where use has been made of the orthonormality of the wave functions:

$$\int \psi^*_{l'm'n'} \psi_{lmn} \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z = \delta_{mm'}\delta_{ll'}\delta_{nn'} \tag{9.11}$$

The matrix elements $M_{lmnl'm'n'}$ are given by:

$$M_{lmnl'm'n'} = -\frac{\hbar^2}{2} \int \psi^*_{l'm'n'} \nabla \left(\frac{1}{m^*(x,y,z)} \nabla \psi_{lmn} \right) \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z$$

$$+ \int \psi^*_{l'm'n'} V \psi_{lmn} \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z \tag{9.12}$$

Carrying out the differentiation, the first integral of equation (9.12) becomes:

$$-\frac{\hbar^2}{2} \left[\int \psi^*_{l'm'n'} \left(\nabla \frac{1}{m^*(x,y,z)} \right) (\nabla \psi_{lmn}) \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z \right.$$

$$\left. + \int \psi^*_{l'm'n'} \frac{1}{m^*(x,y,z)} \nabla (\nabla \psi_{lmn}) \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z \right] \tag{9.13}$$

The second integral of equation (9.13) is then integrated by parts:

$$\int \psi^*_{l'm'n'} \frac{1}{m^*(x,y,z)} \nabla (\nabla \psi_{lmn}) \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z$$

$$= \psi^*_{l'm'n'} \frac{1}{m^*(x,y,z)} \nabla \psi_{lmn} \Big|_{\text{cuboid}} - \int \nabla \left(\psi^*_{l'm'n'} \frac{1}{m^*(x,y,z)} \right) \nabla \psi_{lmn} \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z \tag{9.14}$$

The first term (the non-integral one) of equation (9.14) vanishes (remember that the wave functions vanish at the boundaries of the cuboid), and one is left with:

$$-\int \frac{1}{m^*(x,y,z)} (\nabla \psi^*_{l'm'n'}) (\nabla \psi_{lmn}) \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z$$

$$-\int \psi^*_{l'm'n'} \left(\nabla \frac{1}{m^*(x,y,z)} \right) (\nabla \psi_{lmn}) \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z$$

the second term of which cancels the first integral of equation (9.13), so that only one term is left (besides the one containing the potential in equation (9.12)). One finally gets:

$$M_{lmnl'm'n'} = \frac{\hbar^2}{2}\int \frac{1}{m^*(x,y,z)}\nabla\psi^*_{l'm'n'}\nabla\psi_{lmn}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z + \int \psi^*_{l'm'n'}V\psi_{lmn}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \tag{9.15}$$

The problem here is still the spatial dependence of both the potential and the effective mass in the integrals. While in principle even in the framework of the effective mass approximation such dependence has a complex form (mainly due to the inhomogeneous strain profile), in order to simplify the calculation, V and m^* can be considered constant within each material (a more rigorous justification for this will be given in Section 9.4), i.e. their value exhibits a step-like discontinuity in passing from the barrier region into the dot and another, opposite to the previous, going from the dot into the barrier region. To handle this discontinuity computationally, each integral is split into three parts, within each of which the potential and the effective mass are constant. First, one takes an integral with $m^* = m_\mathrm{B}$ ($V = V_\mathrm{B}$) over the whole cuboid (i.e. barrier plus dot regions). Second, the integral with $m^* = m_\mathrm{B}$ ($V = V_\mathrm{B}$) over the dot region is subtracted. Third, the integral with $m^* = m_\mathrm{D}$ ($V = V_\mathrm{D}$) over the dot region is added:

$$\begin{aligned} M_{lmnl'm'n'} = \frac{\hbar^2}{2}\Bigg[&\frac{1}{m_\mathrm{B}}\int_\mathrm{B}\nabla\psi^*_{l'm'n'}\nabla\psi_{lmn}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z - \frac{1}{m_\mathrm{B}}\int_\mathrm{D}\nabla\psi^*_{l'm'n'}\nabla\psi_{lmn}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \\ &+\frac{1}{m_\mathrm{D}}\int_\mathrm{D}\nabla\psi^*_{l'm'n'}\nabla\psi_{lmn}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\Bigg] + V_\mathrm{B}\int_\mathrm{B}\psi^*_{l'm'n'}\psi_{lmn}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \\ &-V_\mathrm{B}\int_\mathrm{D}\psi^*_{l'm'n'}\psi_{lmn}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z + V_\mathrm{D}\int_\mathrm{D}\psi^*_{l'm'n'}\psi_{lmn}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \end{aligned} \tag{9.16}$$

where the subscripts B and D in the integrals mean that the integration is over the barrier and dot region, respectively. The product in the first integral (I_1) of equation (9.16) gives three similar terms. Consider first the term with the derivative in x:

$$\begin{aligned} I_1^x &= \int_\mathrm{B}\frac{\partial\psi_{l'm'n'}}{\partial x}\cdot\frac{\partial\psi_{lmn}}{\partial x}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \\ &= \frac{2}{L_x}\pi^2\frac{ll'}{L_x^2}\int_{-L_x/2}^{L_x/2}\cos\left[l'\pi\left(\frac{1}{2}-\frac{x}{L_x}\right)\right]\cos\left[l\pi\left(\frac{1}{2}-\frac{x}{L_x}\right)\right]\mathrm{d}x\,\delta_{mm'}\delta_{nn'} \end{aligned}$$

Using the relation $2\cos z_1\cos z_2 = \cos(z_1 - z_2) + \cos(z_1 + z_2)$, the product of the cosines can be expressed as a sum:

$$\begin{aligned} I_1^x = \frac{2}{L_x}\frac{ll'}{L_x^2}\pi^2\delta_{mm'}\delta_{nn'}\frac{1}{2}\int_{-L_x/2}^{L_x/2}\Bigg\{&\cos\left[\pi(l'-l)\left(\frac{1}{2}-\frac{x}{L_x}\right)\right] \\ &+\cos\left[\pi(l'+l)\left(\frac{1}{2}-\frac{x}{L_x}\right)\right]\Bigg\}\,\mathrm{d}x \end{aligned} \tag{9.17}$$

which is easily integrated, giving:

$$I_1^x = \frac{2}{L_x}\frac{ll'}{L_x^2}\pi^2\delta_{mm'}\delta_{nn'}\frac{1}{2}\left\{-\frac{\sin\left[\pi(l'-l)\left(\frac{1}{2}-\frac{x}{L_x}\right)\right]}{\frac{\pi(l'-l)}{L_x}}\Bigg|_{-L_x/2}^{L_x/2}\right.$$
$$\left.-\frac{\sin\left[\pi(l'+l)\left(\frac{1}{2}-\frac{x}{L_x}\right)\right]}{\frac{\pi(l'+l)}{L_x}}\Bigg|_{-L_x/2}^{L_x/2}\right\}. \tag{9.18}$$

The second term in the braces vanishes for all l and l'. The first term, on the other hand, is zero for all $l \neq l'$, i.e. is equal to $\delta_{ll'}$. In fact, it vanishes for all l and l' when evaluated at $L_x/2$ (it behaves like $(\sin x^2)/x$ for $x \to 0$), but it is equal to 1 for $l = l'$, when evaluated at $-L_x/2$ (as the usual limit $(\sin x)/x$ for $x \to 0$). It follows that:

$$I_1^x = \pi^2\frac{ll'}{L_x^2}\delta_{mm'}\delta_{nn'}\delta_{ll'}. \tag{9.19}$$

The same holds for the terms containing the derivatives along y and z. The first integral over the barrier region therefore gives:

$$\int_{\mathrm{B}} \nabla\psi^*_{l'm'n'}\nabla\psi_{lmn}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z = \pi^2\left(\frac{ll'}{L_x^2}+\frac{mm'}{L_y^2}+\frac{nn'}{L_z^2}\right)\delta_{ll'}\delta_{mm'}\delta_{nn'} \tag{9.20}$$

The other integral over the barrier region is simply the normalisation of the wave functions (equation (9.11)), i.e. $\delta_{ll'}\delta_{mm'}\delta_{nn'}$, and the only integrals left in the final expression are those over the dot:

$$M_{lmnl'm'n'} = \left[\frac{\hbar^2\pi^2}{2}\frac{1}{m_{\mathrm{B}}}\left(\frac{ll'}{L_x^2}+\frac{mm'}{L_y^2}+\frac{nn'}{L_z^2}\right)+V\right]\delta_{ll'}\delta_{mm'}\delta_{nn'}$$
$$+\frac{\hbar^2}{2}\left(\frac{1}{m_{\mathrm{D}}}-\frac{1}{m_{\mathrm{B}}}\right)\int_{\mathrm{D}}\nabla\psi^*_{l'm'n'}\nabla\psi_{lmn}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$$
$$-V\int_{\mathrm{D}}\psi^*_{l'm'n'}\psi_{lmn}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \tag{9.21}$$

where the confining potential in the dot region is zero ($V_{\mathrm{D}} = 0$) and $V_{\mathrm{B}} = V$. While the calculation described so far applies to dots with any shape, the evaluation of these two remaining integrals requires a mathematical expression for the dot boundaries. In other words, the shape of the dot needs to be specified at this point.

A relevant feature of this method is that all the integrals in equation (9.21) can be performed analytically. Since the integration limits for two of the three variables (x and y are chosen here) depend on the third one (z, along the height), those two integrations can be

performed independently.

$$\begin{aligned}
&\int_{\mathrm{D}} \psi^*_{l'm'n'}\psi_{lmn}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \\
&= \frac{8}{L_xL_yL_z}\int_z\int_{x(z)} \sin\left[l'\pi\left(\frac{1}{2}-\frac{x}{L_x}\right)\right]\sin\left[l\pi\left(\frac{1}{2}-\frac{x}{L_x}\right)\right]\mathrm{d}x \\
&\quad\times\int_{y(z)} \sin\left[m'\pi\left(\frac{1}{2}-\frac{y}{L_y}\right)\right]\sin\left[m\pi\left(\frac{1}{2}-\frac{y}{L_y}\right)\right]\mathrm{d}y \\
&\quad\times\sin\left[n'\pi\frac{z}{L_z}\right]\sin\left[n\pi\frac{z}{L_z}\right]\mathrm{d}z
\end{aligned} \tag{9.22}$$

The general expressions for the integral of a product of sine functions are (see [17]):

$$\int \sin(ax+b)\sin(cx+d)\,\mathrm{d}x = \frac{\sin\left[(a-c)x+b-d\right]}{2(a-c)} - \frac{\sin\left[(a+c)x+b+d\right]}{2(a+c)}$$

$$\int \sin(ax+b)\sin(ax+d)\,\mathrm{d}x = \frac{x}{2}\cos(b-d) - \frac{\sin(2ax+b+d)}{4a}$$

where the first expression holds for $[a^2 \neq c^2]$.

In the present case $a = -h\pi/L_i$, $b = h\pi/2$, $c = -h'\pi/L_i$ and $d = h'\pi/2$ (where $h = l, m$ and $i = x, y$). Four different situations therefore arise, depending on whether $h = h'$ or not:

(i) $l = l'$ and $m = m'$;

(ii) $l = l'$ and $m \neq m'$;

(iii) $l \neq l'$ and $m = m'$;

(iv) $l \neq l'$ and $m \neq m'$.

Within each of these, there are two more distinct cases, when the integration over z is finally considered:

(a) $n = n'$;

(b) $n \neq n'$.

The situation is similar for the integral containing the product of the gradients of the wave functions.

Having calculated the 16 different analytical expressions for the matrix elements, equation (9.10) can be solved by direct diagonalisation using standard mathematical software such as LAPACK [18].

9.3.1 Convergence tests

Once a specific structure has been identified and the geometry of the problem has been determined (i.e. the shape and composition of both dot and matrix material have been

selected), two important parameters must be chosen: the size of the barrier region (which is also known as the supercell) and the number of basis states that will be used in the calculation (which corresponds to the value of the cut-off wave vector $\mathbf{k}_{\text{cut}}$ in reciprocal space). These quantities, which are mutually interdependent (i.e. the choice of one of them affects that of the other, as, for example, a larger supercell requires a larger number of basis states to achieve a given degree of accuracy in the calculation of the eigenvalues), have the largest influence on the calculation of the electronic structure of any system.

It has been previously mentioned that the boundaries L_x, L_y, L_z must be chosen so that the energy eigenvalues are essentially independent of their choice. The L_i (where $i = x, y, z$) can be defined in terms of the barrier dimensions, i.e. the smallest distance from the dot to the outer box along the three axes, b_x, b_y and b_z, and the dot dimensions l_x, l_y and l_z (where l_x and l_y are the base diameters, and $l_z = h$ is the height), as $L_i = 2b_i + l_i$.

If the convergence tests are performed setting naively $b_x = b_y = b_z$ (a reasonable choice aimed at minimising the number of independent parameters), problems arise for some matrix elements (i.e. their denominator vanishes) whenever one of the conditions $il_{\text{h}}/L_{\text{h}} = jl_k/L_k$ occurs (where i and j are integers and $h, k = x, y, z$). These problems can be overcome by multiplying the barrier dimensions along x, y and z by different constants, c_x, c_y and c_z (all ≈ 1, so that, from the point of view of the energy calculations, the b_i can be considered equal), which avoids the occurrence of the divergence conditions. A set of calculations then has to be performed for each value of the parameters L_i, using different numbers of basis states.

An example of such a calculation for the ground state energy eigenvalue in the conduction band of an InAs square-based pyramid with $b = 120\,\text{Å}$ and $h = 60\,\text{Å}$ embedded in a GaAs matrix is shown in Fig. 9.2, where b_i values range from 25 to 300 Å, and the total number of basis states used is $n_{\text{tot}} = n_{\text{wf}}^3$, as n_{wf} (varying from 7 to 23) has been chosen to be the same along all three directions x, y and z for simplicity (the material parameters used are listed in Table 9.1 below).

As expected, the number of wave functions required to achieve convergence (defined as the achievement of a minimum for the $y = f(x)$ curve, whose value does not change by increasing the value of x; in the present case $E = f(n_{\text{wf}})$), increases with increasing b_i ($n_{\text{wf}} = 15, 19, 19, 21$ for $b_i = 25, 50, 100, 200\,\text{Å}$, respectively). However, from Fig. 9.2 another important feature also emerges: if b_i is too small compared with the dot size (e.g. $b_i = 25\,\text{Å}$), the calculated energy eigenvalues will be too high, no matter how many basis states are used in the expansion of the dot envelope function. In this case the energies are said to be unconverged with respect to the supercell size ($E = f(b_i)$). An indication that convergence in this respect is being achieved is provided by the fact that the difference in the energies calculated with $50 \leq b_i \leq 300$ is only $2\,\text{meV}$, for $n_{\text{wf}} = 23$. (A further effect, not shown in Fig. 9.2, of increasing b_i is the decrease in the spacing between excited states energies.)

The choice $b_i = 200\,\text{Å}$ is found to ensure convergence within $1\,\text{meV}$ for a wide range of pyramid sizes when a basis set of 6859 (i.e. 19^3) wave functions is used (i.e. using $b_i = 200\,\text{Å}$, the difference from the converged energy eigenvalue is only $1\,\text{meV}$, as in the example in Fig. 9.2), while giving the lowest excited states energies, compatible with that number of wave functions and the condition of convergence of the ground state energy. Unlike in Fig. 9.2, where $E(b_i = 100) = E(b_i = 200)$ for $n_{\text{wf}} = 19$, a barrier dimension

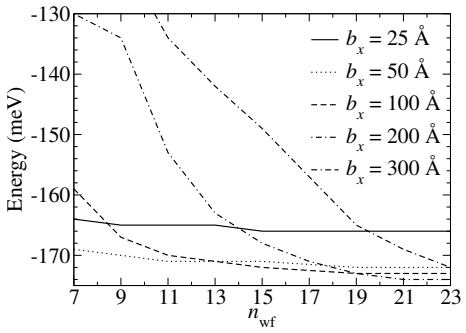


Figure 9.2: Electronic ground state energy levels (with respect to the GaAs conduction band), as a function of the number of wave functions along each direction (i.e. $l, m, n = 1, \ldots, 23$ in equation (9.7): the total number of wave function used is n_{wf}^3, and n_{wf}^3 is the order of the matrix), for an InAs square-based pyramidal quantum dot with $b = 120$ Å and $h = 60$ Å, and different barrier dimensions $b_x (= b_y = b_z)$: $b_x = 25$ Å, dashed line; $b_x = 50$ Å, dot-dashed line; $b_x = 100$ Å, dotted line; $b_x = 200$ Å, solid line; $b_x = 300$ Å, long dashed line

$b_i = 100$ Å proves too small to yield convergence in large pyramids.

In order to improve accuracy further, one would need to increase the value of b_i; however, this comes at the cost of increasing n_{wf}: more than 12 167 (i.e. 23^3) wave functions are needed for $b_i = 300$ Å in order to obtain a similar accuracy ($1\,\text{meV}$), which translates into a memory requirement of $\sim$1 Gb to store the matrix. The final choice of the parameters is therefore determined as a trade-off between accuracy and computational demands. For many practical purposes, an accuracy of $1\,\text{meV}$ is more than adequate, but clearly the memory requirement (and financial cost) of computer hardware increases significantly if further improvements in accuracy are required.

9.3.2 Efficiency

Given the memory constraints just discussed, a legitimate question to ask is whether the choice of basis states (equation (9.7)) was the most efficient. In other words, is it possible, by using a different basis set (still solutions of the infinite cube), to decrease the number of states used in the expansion of the envelope function, retaining at the same time the same level of accuracy? A popular choice in the literature is the use of plane waves $\exp(ik_j x_j)$ (where $j = 1, 2, 3$, x_j are the spatial coordinates x, y and z, $k_j = 2\pi h / L_j$ and $h = l, m, n$). In this basis, the envelope function could be expressed in a compact form as:

$$\begin{aligned}\Psi(x, y, z) &= \sum_{lmn} \frac{a_{lmn}}{\sqrt{L_x L_y L_z}} \exp\left(2\pi i \left(\frac{l}{L_x} + \frac{m}{L_y} + \frac{n}{L_z}\right)\right) \\ &= \sum_{\mathbf{k}} \frac{a_{\mathbf{k}}}{\sqrt{\Omega}} \exp(i\mathbf{k} \cdot \mathbf{r}) \end{aligned} \tag{9.23}$$

The two basis sets are compared here in order to determine their efficiency in terms of:

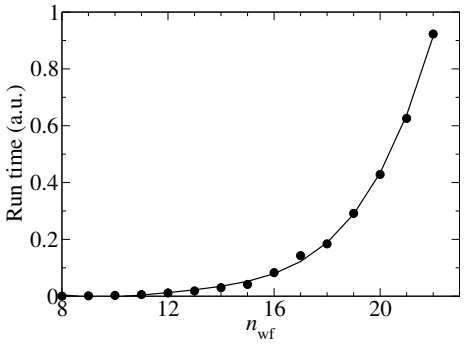


Figure 9.3: Run time as a function of the cube root of the order of the matrix to be diagonalised (i.e. the number of wave functions n_{wf} along each direction). The measured run times have been rescaled to a 0–1 range to preserve generality of results between different computational hardware. The line is a merely a quartic polynomial fitted to guide the eye

- converged energy values, i.e. do both methods give the same set of eigenvalues?
- number of wave functions required to achieve convergence, i.e. how many wave functions are needed in each method to obtain convergence?
- memory requirements, i.e. how much memory is needed to perform the calculations?
- run time, i.e. how long do the calculations take?

The mathematical expressions of the matrix elements for the plane wave calculation are much easier to calculate and more compact to write (they take about one-quarter of the number of lines of code, compared with the sine wave formulation). Nevertheless, when the results of equivalent calculations (i.e. calculations where the same sets of parameters and the same number of wave functions n_{wf} were used) are compared, it is found that the sine wave method gives quicker convergence, in terms of both the number of wave functions employed and the computation time. The latter depends on the order of the matrix and on the size of each element, therefore the more matrix elements to be computed (i.e. the larger n_{wf}), the longer the time required to diagonalise the matrix (see Fig. 9.3).

Also, considering matrices of the same order, the larger the size of the single matrix element (in terms of memory), the longer the diagonalisation time. The matrix elements in the plane wave calculation are *complex* quantities, and therefore occupy twice as much memory as the (*real*) elements of the sine wave method.

Figure 9.4 compares the ground state energies calculated with the two methods, for a pyramid with $b = 120\,\text{Å}$ and $h = 60\,\text{Å}$ (the barrier dimension is $b_i = 200\,\text{Å}$): in this case, the plane wave calculation gives an energy at least $7\,\mathrm{meV}$ larger than that obtained with the sine wave calculation, for all the n_{wf} considered. This means that, in order to obtain the same accuracy, the plane wave calculation requires more wave functions (i.e. bigger matrices) and takes much longer to run than the sine wave method. This proves that sine waves are, from all points of view, a more efficient basis set than plane waves for the expansion of the envelope function when performing such calculations.

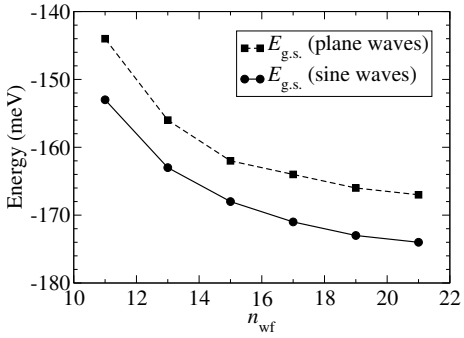


Figure 9.4: Electronic ground state energy levels as a function of the equivalent number of wave functions along each direction (the total number of wave function used is n^3), for an InAs square-based pyramidal quantum dot with $b = 120\,\text{Å}$ and $h = 600\,\text{Å}$, with respect to the GaAs conduction band. Solid line, sine wave calculations; dashed line, plane wave calculations

It is perhaps worth mentioning that, although Barker *et al.* [11] claim that only 11 plane waves along the in-plane directions and 17 along the growth direction are enough to ensure convergence, for a pyramid of the dimensions considered here, they chose a separation between neighbouring dots equal to the dot dimension, which means a barrier dimension of half this quantity (i.e. $b = 60\,\text{Å}$). As already discussed above for the sine wave calculation, this choice ensures a quicker convergence (i.e. fewer wave functions are required to achieve it), but the converged eigenvalue is higher than that obtained when larger barriers are used (as seen in Fig. 9.2). In order to confirm this, another set of plane wave calculations is performed, using a barrier dimension of $60\,\text{Å}$: the converged ground state energy (obtained with $n_{\text{wf}} = 19$ along all directions) is found to be $E_{\text{g.s.}} = -170\,\text{meV}$, which is only $1\,\text{meV}$ lower than the value obtained with 11 plane waves along the in-plane directions and 17 along the growth direction, but it is $5\,\text{meV}$ higher than the value found with the sine wave calculation and a barrier of $200\,\text{Å}$, and $2\,\text{meV}$ higher than the energy obtained with the sine wave method and the same barrier of $60\,\text{Å}$.

9.3.3 Optimisation

Having proved the superior efficiency of the sine wave formulation compared with the full plane wave approach, it would now be worthwhile to explore possible simple strategies to optimise the code performance, exploiting the distinctive characteristics of the specific system treated. Figure 9.3 shows the calculation run time as a function of the cube root of the order of the matrix ($n_{\text{wf}}^3 \times n_{\text{wf}}^3$), or, in other words, the number of wave functions n_{wf} along each direction (where $l, m, n = 1, \ldots, n_{\text{wf}}$ in equation (9.7); so far, the same n_{wf} has been used along the three directions x, y, z, in the expansion of the envelope function).

A simple general strategy that can be adopted in order to reduce the calculation time is the reduction of the *accuracy* of the quantities used in the calculations from *double* precision to *single* precision (i.e. `double`→`float` in C, `real*8` → `real*4` in Fortran), reducing

their size from 8 to 4 bytes. This means that the matrix requires less memory to store (it occupies a quarter of the memory needed by a matrix made of doubles), and less time (about 25% less) to diagonalise. The accuracy of the eigenvalues, however, is not affected in this case by this change of computational accuracy.

Another improvement in terms of run time can be obtained by exploiting the symmetries of the system. In a pyramid, for example, the height is the symmetry axis, therefore the potential is symmetric along x and y, and the in-plane (xy) wave functions can be of either even or odd parity with respect to that axis. If, in the calculations only even (odd) parity functions are considered along x and y, and all even- and odd-parity functions along z (to take into account that the potential is not symmetric along that direction), one might expect to obtain eigenvalues and eigenvectors of only even (odd) parity states. Unfortunately, this conclusion proves not to be completely correct: the calculations with even-parity wave functions yield the energy of the ground state (as expected) and of some excited states (fourth, fifth, sixth, etc.), whereas the use of odd-parity functions produces the third excited state energy and some higher ones. No calculation is, however, able to reproduce the energies of the first two excited states, which implies that they must contain a superposition of both even- and odd-parity wave functions. Nevertheless, if only the ground state energy is needed, this method proves to be very efficient and reliable, with very low computational demands: it reproduces the results obtained with $n_{\mathrm{wf}} = 19$ in the full calculation, using only 10 wave functions along x and y and 19 along z (i.e. a matrix 13 times smaller), in less than 4% of the run time of the unoptimised code.

9.4 Application to real problems

It is now instructive, from the point of view of the aims of this book, to give an example of how such a very simple computational method can be developed into a simple physical model, which yields good agreement with experimental results. To show its potential, it will be applied to the study of the electronic and optical properties of InAs self-assembled quantum dots in a GaAs matrix and its results compared both with the predictions of much more complex theoretical approaches and with experimental data relative to a few different samples.

9.4.1 InAs/GaAs self-assembled quantum dots

As discussed in Chapter 7 for the more general case of a quantum well, the difference in lattice constant between InAs and GaAs (where $a_{\mathrm{InAs}} > a_{\mathrm{GaAs}}$), which provides the key ingredient for the formation of the dots, leads to the build-up of a strain field in the system, which affects both the confining potential and the effective mass of the carriers. The confining potential becomes a piecewise continuous function of position and differs from the square well formed by the difference in the absolute energy of the conduction- or valence-band edges in the bulk dot and barrier material. In the dot, the compressive stress alters the curvature of the bulk bands, causing also the effective masses to differ from the unstrained ones. The presence of shear strain, moreover, leads to the appearance of polarisation charges with an associated piezoelectric potential (see Sections 7.6–7.8). Finally, the charge carriers

interact via the Coulomb potential, giving rise to additional terms in the Hamiltonian (see equation (6.7)).

9.4.2 Working assumptions

This complex picture seems to undermine the simplicity of the computational method illustrated in Section 9.3. However, the complexity of the problem can be considerably reduced, without compromising the integrity of the physical description of the system, by introducing the following approximations:

(i) the motion of electrons and holes is decoupled;

(ii) the effects of strain on both the confining potentials and effective masses are accounted for *on average*;

(iii) piezoelectric effects and Coulombic interactions are neglected.

In this framework (the physical model), therefore, the method can be easily applied, as: (i) the energy levels of electrons and holes are calculated with separate one-band Hamiltonians within the envelope function effective mass approximation; (ii) the confining potentials and effective masses are assumed to be constant (and equal to their average *strained* value) throughout the dot and the matrix for both carriers and for all dot sizes (see Table 9.1); (iii)(a) piezoelectric effects are disregarded, as in single dots of realistic sizes they generally affect the energies of the levels involved in optical transitions by less than $1\,\mathrm{meV}$ [19] (although the effects may be considerably larger in systems of closely spaced dots, since their amplitude is different in each dot: theoretical calculations show [20] that piezoelectricity is one of the most important factors contributing to the level splitting in stacked quantum dots); (iii)(b) the Coulombic interaction is also neglected, since the quantum dots considered are in the *strong confinement regime* [21], where the size quantisation represents the main part of the carrier energy (their effective radius $r = b/(8\pi)^{1/3}$, where b is the pyramid base, is small compared with the bulk exciton Bohr radius $a_{0,\mathrm{InAs}} = 340\,\text{Å}$).

Table 9.1 Calculation parameters: m_B barrier region effective mass; m_D dot region effective mass (all in units of the bare electron mass m_0); V_0 carrier confining potential (in millielectronvolts).

Electron			Heavy hole		
m_B	m_D	V_0	m_B	m_D	V_0
0.0665^a	0.040^b	450^a	0.3774^a	0.341^a	316^c

$^{(a)}$ [19], $^{(b)}$ [22], $^{(c)}$ [12]

9.4.3 Results

The ground state energy eigenvalues for both electrons and heavy holes, calculated using this simple model, are presented in Fig. 9.5 and compared with the results of more sophisticated theoretical approaches [14, 19] which take into account band mixing, piezoelectric effects and the spatial variation of the confining potentials due to the strain distribution. The agreement is very good for both the conduction and the valence band, especially considering the complex nature of the latter.

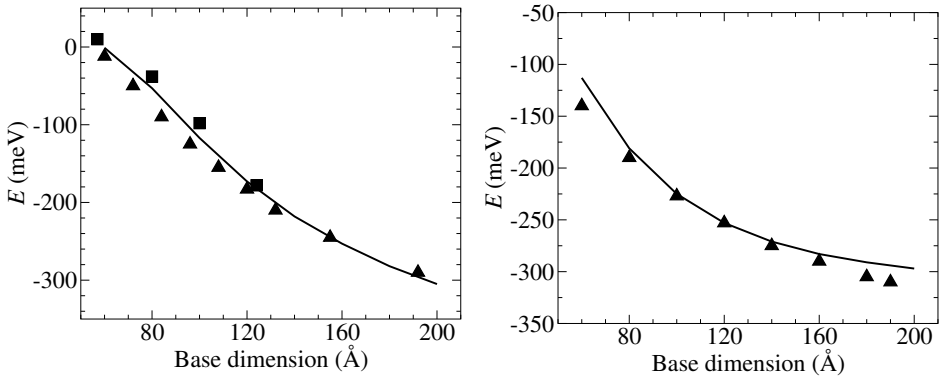


Figure 9.5: Electron (left) and heavy-hole (right) ground state energy levels as a function of the base dimension, for InAs square based pyramidal quantum dots with aspect ratio 1, with respect to the unstrained GaAs conduction and valence bands, respectively. Comparison of the results of the present model (solid lines) with those of [14] (squares) and [19] (triangles)

The model predicts no bound electron states for base lengths smaller than about $60\,\text{Å}$, one bound state for $60\,\text{Å} < b < 100\,\text{Å}$ and up to ten for $b = 200\,\text{Å}$. Interestingly, the supposedly more accurate (and certainly more elaborate) approach of [19] predicts only one electron state for this size. The importance of this feature will appear more clearly later on when the results of the simple model are compared with experimental data. The prediction of three bound states in quantum dots with $b = 120\,\text{Å}$, five for $b = 160\,\text{Å}$ and seven for $b = 180\,\text{Å}$ is in excellent agreement with calculations reported by Stier *et al*. [23] who found 4 and 6 bound electron states in structures with $b = 136\,\text{Å}$ and $170\,\text{Å}$ respectively, by modelling strained quantum dots with the much more sophisticated eight-band $\mathbf{k}.\mathbf{p}$ theory. As in [14], the first and second excited electron states are found to be degenerate, as expected, due to the C_{4v} symmetry of the pyramidal dot. Figure 9.6 shows the electronic wave functions $\Psi(x, y, z)$ (calculated at the base of the dot), for the ground and the first excited state (higher excited states are not bound), in an InAs pyramid with $b = 120\,\text{Å}$ and $h = 60\,\text{Å}$.

The predictions of the model are next compared with several photoluminescence (PL) spectra, relative to samples with different sizes [24, 25, 26, 27], in Tables 9.2, 9.3 and 9.4 (and in the text), where the identification of the peaks, in terms of transitions between specific electron and hole single-particle states, follows the attribution proposed in the original papers.

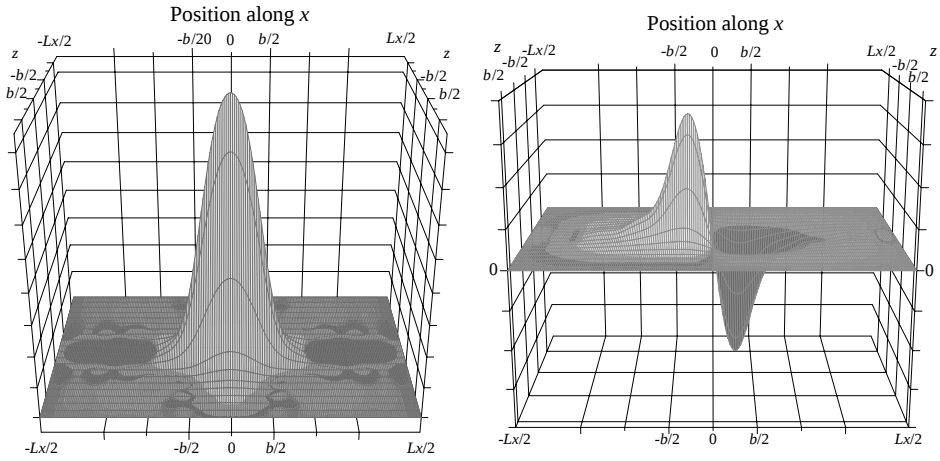


Figure 9.6: Electronic wave function Ψ (calculated at the base of the dot), for the two lowest-energy conduction-band states in an InAs pyramid with $b = 120\,\text{Å}$ and $h = 60\,\text{Å}$

Table 9.2 Transition energies: comparison between experimental data (deduced from PL spectra reported in [25]), and theoretical results for a square-based pyramid with $b = 110\,\text{Å}$ and $h = 60\,\text{Å}$. Since the model predicts the first and second excited hole states to be degenerate, the third peak has been attributed to a transition to the third excited hole level

Transition	Experimental, $E_{\text{exp.}}$ (eV)	Theoretical, $E_{\text{th.}}$ (eV)	Error, $\lvert\delta E\rvert$ (eV)
$e_0 \to h_0$	1.10	1.112	0.012
$e_0 \to h_{1(2)}$	1.17	1.176	0.006
$e_0 \to h_3$	1.24	1.224	0.016

The emission energy for the transition between states e_i in the conduction band and h_j in the valence band are computed as:

$$E(e_i \to h_j) = E_{\text{g}}(\text{GaAs}) + E(e_i) + E(h_j)$$

where $E_{\text{g}}(\text{GaAs})$ is the energy gap of bulk GaAs ($1.519\,\text{eV}$) and $E(e_i), E(h_j)$ are respectively the electron and hole energies relative to the unstrained GaAs band edges (which, as in the example in Fig. 9.5, assume negative values if relative to bound states).

The agreement with both the number and the position of the observed spectral peaks over a wide range of dot sizes and aspect ratios is very good, especially considering the simplicity of the model: the error δE between calculated and measured energies is always much smaller than the broadening of the PL peaks whose full width at half maximum typical values are of the order of 60–70 meV.

It is interesting to note that, for an InAs pyramidal dot with $b = 200\,\text{Å}$ and $h = 70\,\text{Å}$, the simple model predicts five different electron states (actually six bound states are predicted, the first and second excited states being degenerate so that there are only five distinct energy

Table 9.3 Transition energies: comparison between experimental data (deduced from PL spectra of [24]), and theoretical results for an InAs pyramidal dot with $b = 200$ Å and $h = 70$ Å

Transition	Experimental, $E_{\text{exp.}}$ (eV)	Theoretical, $E_{\text{th.}}$ (eV)	Error, $\lvert\delta E\rvert$ (eV)
$e_0 \rightarrow h_0$	1.01	0.979	0.022
$e_{1,2} \rightarrow h_{1,2}$	1.10	1.112	0.012
$e_3 \rightarrow h_3$	1.17	1.214	0.044
$e_4 \rightarrow h_4$	1.23	1.237	0.007
$e_5 \rightarrow h_5$	1.29	1.264	0.026

levels), whereas the more complex approaches of [14, 19, 28] are unable to reproduce this fundamental feature. The importance of this result becomes clear when considering that the PL spectrum of such a sample grown by Schmidt *et al*. [24] shows precisely five peaks (see Table 9.3), which were attributed to transitions between electron and hole states with the same quantum numbers and are therefore consistent with the existence of five different electronic energy levels in the quantum dot. Furthermore, the double degeneracy of the first electron and hole excited states predicted by the simple model is also mentioned in [24]. The calculated energy splitting $\Delta E^{\text{th.}}_{h0,h1} = 29\,\text{meV}$ between the ground and first excited heavy-hole states in this structure is in excellent agreement with the experimentally estimated value of $\Delta E^{\text{exp.}}_{h0,h1} = 27\,\text{meV}$ obtained by combining capacitance and PL measurements [24]. The predicted electron energy splitting $\Delta E^{\text{th.}}_{e0,e1}$ however, is about 50% larger ($104\,\text{meV}$) than the experimental value deduced from the separation between the first two PL peaks (i.e. $\Delta E_{e0,e1} \approx E(e_1 \rightarrow h_1) - E(e_0 \rightarrow h_0) - \Delta E_{h0,h1} = 63\,\text{meV}$).

These results are surprising, considering that the constant confining potential approximation used here is expected to give a better description of the conduction than the valence band [19]. In fact, whereas the conduction-band offset is almost constant throughout the dot, the use of such a simplistic assumption in the valence band is expected to affect the hole level alignment because of the more complex shape of the confinement's real profile in structures with an aspect ratio of 1 (see [12, 29]). This approximation should be more suitable for flatter structures with high aspect ratios, where the confining potentials for both the electrons and holes have almost a square well shape [30]. As discussed in Section 9.4.4 below, the suggestion by Nishiguchi and Yoh [31], of including an energy dependence for the effective mass to reduce the overestimate of the energy level separation in the conduction band, would not improve the agreement with the experimental data.

For a rectangular based pyramidal quantum dot of height 30 Å and base dimensions 250 and 300 Å along the [110] and [1$\bar{1}$0] directions, respectively, the model predicts six bound electron states (i.e. the ground state and five excited states), thus the absence of any value for the transition between the sixth excited electron and heavy-hole energy levels in Table 9.4. However, Noda *et al*. [26] mention that this last peak may be due to the wetting layer signal as well.

Another spectral position reproduced well by the model is that obtained by Toda *et al*. [27] in near-field magneto-optical spectroscopy measurements of single self-assembled quantum dots. The structures they investigate have lateral size ~ 200 Å and height ~ 20 Å, as indicated by atomic force microscope studies of uncapped layers. The typical magnetic field dependence of the peak energies from a single quantum dot they show has a value

Table 9.4 Transition energies: comparison between experimental data (deduced from PL and PL excitation spectra of [26]), and theoretical results for a rectangular-based pyramidal quantum dot of height 30 Å and base dimensions 250 and 300 Å along the [110] and [1$\bar{1}$0] directions, respectively. In this case the model predicts the first and second excited levels not to be degenerate, because of the different symmetry of the system (i.e. the pyramid is not square-based)

Transition	Experimental, $E_{\text{exp.}}$ (eV)	Theoretical, $E_{\text{th.}}$ (eV)	Error, $\lvert\delta E\rvert$ (eV)
$e_0 \rightarrow h_0$	1.220	1.168	0.052
$e_1 \rightarrow h_1$	1.270	1.261	0.009
$e_2 \rightarrow h_2$	1.284	1.280	0.004
$e_3 \rightarrow h_3$	1.332	1.338	0.006
$e_4 \rightarrow h_4$	1.340	1.341	0.001
$e_5 \rightarrow h_5$	1.380	1.378	0.002
$e_6 \rightarrow h_6$	1.412	–	–

of about $1347\,\text{meV}$ for zero magnetic field. The value calculated with the simple model is $1321\,\text{meV}$. Moreover, only one bound electron state is predicted for such structures.

9.4.4 Concluding remarks

It is perhaps worth stressing once more that, although the Schrödinger equation used in the present approach does not explicitly include the strain, its effects are nevertheless accounted for by the choice of average strained values for effective masses and confining potentials (for both electrons and heavy holes). The choice of these parameters is based on a set of calculations of the strain distribution in pyramidal structures with aspect ratio Q ranging from 1 to 4.5 performed by Califano and Harrison [12], using a method based on the Green's function technique [32], where the anisotropy of the elastic properties was taken into account as well. The carriers' strained confining potentials were calculated as a function of position along the growth direction, in the framework of eight-band $\mathbf{k}.\mathbf{p}$ theory. The results showed that the confining potentials of both electrons and holes are almost constant along the dot axis for $Q \geq 2$ (in agreement with previously published data [30]), with values in the centre of the island ranging from about 414 to about $463\,\text{meV}$ (for the electrons), and from 253 to $332\,\text{meV}$ (for the heavy holes), for the experimental structures considered above, proving the constant-confining-potential approximation as a reasonable choice.

Nevertheless, the high degree of agreement obtained by this simple model with the transition energies of so many different experimental samples might appear hard to explain. It seems to strongly indicate that the *details* of the strain distribution are not as important for the transition energy in InAs quantum dots as they are commonly believed to be.

Califano and Harrison [12] found that the strain distribution for a pyramid depends very weakly on the dot volume (for a given aspect ratio Q), but is more sensitive to a change in Q. This variation, however, principally affects the biaxial component (and is more accentuated near the tip of the dot), whereas the hydrostatic strain is only slightly modified. The strain-induced shift of both conduction- and *average* valence-band energies, however, is due only

to the hydrostatic component, i.e.

$$\Delta E_{\mathrm{c}} = a_{\mathrm{c}}\epsilon_{\mathrm{h}} \tag{9.24}$$

$$\Delta E_{\mathrm{hh}} = a_{\mathrm{v}}\epsilon_{\mathrm{h}} - \frac{b}{2}\epsilon_{\mathrm{b}} \tag{9.25}$$

where ΔE_{c}, ΔE_{hh}, ϵ_{h} and ϵ_{b} are the energy shifts of conduction and heavy-hole bands, and the hydrostatic and biaxial components of the strain, respectively (a_{c}, a_{v} and b are the deformation potentials). Therefore, only the biaxial-component-induced heavy-hole energy shift is substantially affected by the variation of the dot aspect ratio. However, while this effect is most pronounced towards the pyramid tip, the main contribution to the integral in the matrix element of the heavy-hole confining potential that appears in the Hamiltonian comes from the bottom of the pyramid, where the heavy-hole wave function is localised (see, for example, [11, 19, 23]), and depends on the pyramid size. Therefore, it is the size quantisation together with the (less size-sensitive) hydrostatic strain distribution (which determines the shift of the *average* band position) that represents the dominating part of the heavy-hole energy. This could explain why the transition energies calculated using the same values for the *average* strained hole confining potential agree so well with experiment for such a wide range of dot sizes, and the excellent agreement obtained for the heavy-hole energy splitting $\Delta E_{h0,h1}$.

The situation is a bit more complex for the electron effective mass tensor, the expression for which depends on the particular approximation adopted. Generally, one can say that, because it is a function of the carriers' band-edge energies, a similar argument applies as for the heavy-hole band edge, except that in this case the energy shifts of light-hole and spin–orbit split-off bands (if their important contribution is included into the effective mass expression) also depend on the biaxial strain. Due to the more complex dependence on the biaxial component of strain, however, the resultant effect of a variation in the aspect ratio is less predictable. Furthermore, in the case of the electron the wave function is localised above that of the heavy holes, yielding a higher value for the matrix element in the region where the difference (of the strain distribution between dots with different values of the aspect ratio) is bigger.

In this case, a compensation mechanism in which a positive difference in the confining potential (i.e. between the assumed average strained value of $450\,\mathrm{meV}$ and the average strained value calculated for the specific sample) is compensated by a negative difference in the value of the effective mass (so that when the confining potential is less than $450\,\mathrm{meV}$, the effective mass is greater than 0.04), and vice versa, could be responsible for the agreement in the transition energies.

Finally, a few words on excitonic effects are perhaps appropriate here as well. The simple model discussed so far calculates the total energy of the electron–hole system (the exciton) as the sum of its *single-particle* energies, i.e. the energy levels of electrons and holes are considered to be independent of the presence of the other particle as if, instead of 'living together' under the same 'pyramidal roof', they were infinitely far apart. However, as in reality the electron and the hole are confined within a space that is much smaller than the exciton Bohr radius in the bulk (the radius of an exciton in bulk InAs is $34\,\mathrm{nm}$), there is considerable Coulombic (i.e. electrostatic) interaction between them due to their opposite charge. This effect can be taken into account perturbatively, and the exciton energy can be

written within first order as:

$$E(h_i, e_j) = E_{\mathrm{g}}^{\mathrm{bulk}} + E(e_j) + E(h_i) - (J_{e_j,h_i} - K_{e_j,h_i}) \tag{9.26}$$

where the first term contains the energies of the (non-interacting) electron and hole, respectively, in levels j and i, and the last term represents direct and exchange Coulombic energies given by:

$$J_{i,j} = \iint \frac{|\psi_i(\mathrm{r}_1)|^2 |\psi_j(\mathrm{r}_2)|^2}{\epsilon(\mathrm{r}_1, \mathrm{r}_2)|\mathrm{r}_1 - \mathrm{r}_2|} \mathrm{d}\mathrm{r}_1 \mathrm{d}\mathrm{r}_2 \tag{9.27}$$

and

$$K_{i,j} = \iint \frac{\psi_i^*(\mathrm{r}_1)\psi_j^*(\mathrm{r}_1)\psi_i(\mathrm{r}_2)\psi_j(\mathrm{r}_2)}{\epsilon(\mathrm{r}_1, \mathrm{r}_2)|\mathrm{r}_1 - \mathrm{r}_2|} \mathrm{d}\mathrm{r}_1 \mathrm{d}\mathrm{r}_2 \tag{9.28}$$

where ϵ is the dielectric constant of the dot. The largest contribution in the second term comes from the direct Coulombic interaction and was calculated to be of the order of a few tens of millielectronvolts [33] for a wide range of dot shapes and has therefore been neglected for simplicity in the present approach. (Incidentally this is also the order of magnitude of the discrepancies between theory and experiment found above; see Tables 9.2–9.4.) The exchange energy, was found [33] to be an order of magnitude smaller than the direct Coulombic term and was similarly neglected. Other many-body effects such as correlations, even though responsible for the excitonic 'fine structure' [34], are not relevant at the level of resolution discussed above, as they are expected to have even smaller contributions to the total exciton energy.

9.5 A more complex model is not always a better model

This simple model can be modified to include a more detailed account of the strain-induced spatial variation of the crucial parameters V and $m_{\mathrm{eff.}}$. As mentioned in the previous section, the strain distribution in self-assembled pyramidal $\mathrm{In}_{1-x}\mathrm{Ga}_x\mathrm{As/GaAs}$ quantum dots can be calculated using a method based on the Green's function technique [32], accounting for the dependence of the biaxial and hydrostatic components on the quantum dot volume, aspect ratio and composition. The position dependence of the strained carriers' confining potentials and electron effective mass (in-plane and perpendicular components) can then be calculated in the framework of eight-band $\mathbf{k}.\mathbf{p}$ theory [12]. In this respect, the inclusion of the coupling with the spin–orbit split-off band in the calculation of the effective mass tensor has been shown [12] to be crucial for a correct description of the effects of the strain.

The spatial dependence of V and $m_{\mathrm{eff.}}$, however, makes it impossible to keep an analytical form for the matrix elements in the Hamiltonian, and resort to a time-consuming numerical integration is needed in the calculation of the electronic energy levels. Unfortunately, this considerable increase in the complexity of the model does not lead to a comparable increase in the accuracy of its results. Calculations performed within the framework of the effective mass approximation [29] showed that the inclusion of such an accurate account of the effect of strain, resulting in position-dependent parameters, does not induce any appreciable variation, at least in the ground state electronic energies, compared with the constant potential and effective mass choice detailed in Table 9.1.

This result seems to confirm the suggestion made above that *details* of the strain distribution are apparently not as important for the transition energies in InAs quantum dots as

they are commonly believed to be, and that simple models that correctly capture the physics of the system can successfully predict its energies too.

Exercises

(1). Explore the sensitivity of the electron ground state energy to the shape of a quantum dot by plotting the energy, within a cuboidal GaAs quantum dot surrounded by infinitely high barriers, as a function of the length along the x-direction *for fixed dot volume*. Hint: restrict changes of the dot shape to two dimensions only.

(2). Building on the previous exercise or otherwise, prove *analytically* that the ground state energy of a cuboidal quantum dot is at a minimum when the dot is a cube.

(3). Given that InAs forms a zinc blende type crystal structure with a lattice constant of 6.06 Å, derive a formula for the number of atoms in a square-based pyramidal quantum dot of base length b and height h. Hence, plot the number of atoms as a function of b for aspect ratios (h/b) of 0.5, 0.75, 1.0, 1.25 and 1.5.

(4). Prove that the basis functions:

$$\psi_{lmn} = \sqrt{\frac{2}{L_x}} \sin\left(\frac{l\pi x}{L_x}\right) \sqrt{\frac{2}{L_y}} \sin\left(\frac{m\pi y}{L_y}\right) \sqrt{\frac{2}{L_z}} \sin\left(\frac{n\pi z}{L_z}\right)$$

are orthonormal across the cuboidal domains $[0, L_x]$, $[0, L_y]$ and $[0, L_z]$.

(5). Take the simplified case of a cuboidal quantum dot surrounded by infinitely high barriers with a constant effective mass, a basis function of which would be a product of orthogonal sine waves:

$$\psi_{lmn} = \sqrt{\frac{2}{L_x}} \sin\left(\frac{l\pi x}{L_x}\right) \sqrt{\frac{2}{L_y}} \sin\left(\frac{m\pi y}{L_y}\right) \sqrt{\frac{2}{L_z}} \sin\left(\frac{n\pi z}{L_z}\right)$$

The wave function of the quantum dot could therefore be expressed as a linear combination of these basis states, i.e.

$$\Psi(x, y, z) = \sum_{lmn} a_{lmn} \psi_{lmn}$$

where, for convenience, the spatial origin has been placed at a corner of the cuboid. The Schrödinger equation follows as $\mathcal{H}\Psi = E\Psi$, which, when multiplied on the left by $\psi^*_{l'm'n'}$ and integrated over all space, leads to the matrix equation:

$$M_{l'm'n',lmn} = E\delta_{l'l}\delta_{m'm}\delta_{n'n}$$

Derive an expression for the matrix elements on the left-hand side of the above equation. Compare the results with the analytical solutions from exercise (1).

(6). Building on the previous example, now define a finite potential energy V within the confines of a cubic dot surrounded by infinite barriers, in order to describe a finite

barrier cuboidal quantum dot potential. Add the potential term to the Hamiltonian $\mathcal{H}$ and hence derive expressions for the corresponding matrix elements.

Write a computer program to perform this three-dimensional finite-barrier (constant effective mass) quantum dot calculation. Experiment with the number of basis states necessary to converge the energy eigenvalues to $1\,\mathrm{meV}$. Hence, compare the results of calculations of electron ground state energies in cubic GaAs quantum dots of various sizes and barrier heights with the analytical solutions of similar dots but surrounded by infinitely high barriers.

References

[1] A. Kumar, S. E. Laux, and F. Stern, 'Electron states in a GaAs quantum dot in a magnetic field', *Phys. Rev. B*, **42**:5166–5175, 1990.

[2] D. J. Eaglesham and M. Cerullo, 'Dislocation-free Stranski–Krastanow growth of Ge on Si(100)', *Phys. Rev. Lett.*, **64**:1943–1946, 1990.

[3] H. Sunamura, N. Usami, Y. Shiraki, and S. Fukatsu, 'Island formation during growth of Ge on Si(100): a study using photoluminescence spectroscopy', *Appl. Phys. Lett.*, **66**:3024–3026, 1995.

[4] M. Grundmann, J. Christen, N. N. Ledentsov, J. Böhrer, D. Bimberg, S. S. Ruvimov, P. Werner, U. Richter, U. Gösele, J. Heydenreich, V. M. Ustinov, A. Y. Egorov, A. E. Zhukov, P. S. Kop'ev, and Z. I. Alferov, 'Ultranarrow luminescence lines from single quantum dots', *Phys. Rev. Lett.*, **74**:4043–4046, 1995.

[5] J. M. Moison, F. Houzay, F. Barthe, L. Leprince, E. André, and O. Vatel, 'Self-organized growth of regular nanometer-scale InAs dots on GaAs', *Appl. Phys. Lett.*, **64**:196–198, 1994.

[6] D. Leonard, K. Pond, and P. M. Petroff, 'Critical layer thickness for self-assembled InAs islands on GaAs', *Phys. Rev. B*, **50**:11687–11692, 1994.

[7] M. Fricke, A. Lorke, J. Kotthaus, G. Medeiros-Ribeiro, and P. Petroff, 'Shell structure and electron–electron interaction in self-assembled InAs quantum dots', *Europhys. Lett.*, **36**:197–202, 1996.

[8] S. Sauvage, P. Boucaud, F. H. Julien, J.-M. Gérard, and J.-Y. Marzin, 'Infrared spectroscopy of intraband transitions in self-organized InAs/GaAs quantum dots', *J. Appl. Phys.*, **82**:3396–3401, 1997.

[9] N. Liu, J. Tersoff, O. Baklenov, A. L. Holmes, and C. K. Shih, 'Nonuniform composition profile in $In_{0.5}Ga_{0.5}As$ alloy quantum dots', *Phys. Rev. Lett.*, **84**:334–337, 2000.

[10] D. Bimberg, M. Grundmann, and N. N. Ledentsov, *Quantum Dot Heterostructures*, John Wiley & Sons, Ltd., Chichester, 1999.

[11] J. A. Barker and E. P. O'Reilly, 'Theoretical analysis of electron–hole alignment in InAs–GaAs quantum dots', *Phys. Rev. B*, **61**(20):13840–13851, 2000.

[12] M. Califano and P. Harrison, 'Presentation and experimental validation of a single-band, constant-potential model for self-assembled InAs/GaAs quantum dots', *Phys. Rev. B*, **61**(16):10959–10965, 2000.

[13] L. R. C. Fonseca, J. L. Jimenez, J. P. Leburton, and R. M. Martin, 'Self-consistent calculation of the electronic structure and electron–electron interaction in self-assembled InAs-GaAs quantum dot structures', *Phys. Rev. B*, **57**:4017–4026, 1998.

[14] M. A. Cusack, P. R. Briddon, and M. Jaros, 'Electronic structure of InAs/GaAs self-assembled quantum dots', *Phys. Rev. B*, **54**:R2300–R2303, 1996.

[15] D. Gershoni, H. Temkin, G. J. Dolan, J. Dunsmuir, S. N. G. Chu, and M. B. Panish, 'Effects of two-dimensional confinement on the optical properties of InGaAs/InP quantum wire structures', *Appl. Phys. Lett.*, **53**(11):995–997, 1988.

[16] S. Gangopadhyay and B. R. Nag, 'Energy levels in three-dimensional quantum-confinement structures', *Nanotechnology*, **8**(1):14, 1997.

[17] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, London, Sixth edition, 2000.

[18] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, Third edition, 1999.

[19] M. Grundmann, O. Stier, and D. Bimberg, 'InAs/GaAs pyramidal quantum dots: Strain distribution, optical phonons, and electronic structure', *Phys. Rev. B*, **52**:11969–11981, 1995.

[20] L. R. C. Fonseca, J. L. Jimenez, and J. P. Leburton, 'Electronic coupling in InAs/GaAs self-assembled stacked double-quantum-dot systems', *Phys. Rev. B*, **58**:9955–9960, 1998.

[21] A. Efros and A. Efros, 'Interband absorption of light in a semiconductor sphere', *Sov. Phys. Semicond.*, **16**:772–775, 1982.

[22] D. Ninno, M. A. Gell, and M. Jaros, 'Electronic structure and optical transitions in GaAs–$Ga_{1-x}Al_xAs$(001) superlattices', *J. Phys. C: Solid State Phys.*, **19**(20):3845, 1986.

[23] O. Stier, M. Grundmann, and D. Bimberg, 'Electronic and optical properties of strained quantum dots modeled by 8-band k·p theory', *Phys. Rev. B*, **59**:5688–5701, 1999.

[24] K. H. Schmidt, G. Medeiros-Ribeiro, M. Oestreich, P. M. Petroff, and G. H. Döhler, 'Carrier relaxation and electronic structure in InAs self-assembled quantum dots', *Phys. Rev. B*, **54**:11346–11353, 1996.

[25] M. Grundmann, N. N. Ledentsov, O. Stier, D. Bimberg, V. M. Ustinov, P. S. Kop'ev, and Z. I. Alferov, 'Excited states in self-organized InAs/GaAs quantum dots: theory and experiment', *Appl. Phys. Lett.*, **68**:979–981, 1996.

[26] S. Noda, T. Abe, and M. Tamura, 'Mode assignment of excited states in self-assembled InAs/GaAs quantum dots', *Phys. Rev. B*, **58**:7181–7187, 1998.

[27] Y. Toda, S. Shinomori, K. Suzuki, and Y. Arakawa, 'Near-field magneto-optical spectroscopy of single self-assembled InAs quantum dots', *Appl. Phys. Lett.*, **73**:517–519, 1998.

[28] J. Marzin and G. Bastard, 'Calculation of the energy-levels in InAs/GaAs quantum dots', *Solid State Commun.*, **92**:437–442, 1994.

[29] M. Califano, *Development of Computational Models for the Electronic Structure of Self-Assembled Quantum Dots*, Ph.D. thesis, University of Leeds, 2002.

[30] M. A. Cusack, P. R. Briddon, and M. Jaros, 'Absorption spectra and optical transitions in InAs/GaAs self-assembled quantum dots', *Phys. Rev. B*, **56**:4047–4050, 1997.

[31] N. Nishiguchi and K. Yoh, 'Energy-dependent effective mass approximation in one-dimensional quantum dots', *Jpn. J. Appl. Phys.*, **36**:3928–3931, 1997.

[32] A. D. Andreev, J. R. Downes, D. A. Faux, and E. P. O'Reilly, 'Strain distributions in quantum dots of arbitrary shape', *J. Appl. Phys.*, **86**:297–305, 1999.

[33] A. J. Williamson, L. W. Wang, and A. Zunger, 'Theoretical interpretation of the experimental electronic structure of lens-shaped self-assembled InAs/GaAs quantum dots', *Phys. Rev. B*, **62**:12963–12977, 2000.

[34] A. Franceschetti, H. Fu, L. W. Wang, and A. Zunger, 'Many-body pseudopotential theory of excitons in InP and CdSe quantum dots', *Phys. Rev. B*, **60**:1819–1829, 1999.

10

Carrier scattering

10.1 Introduction

So far, the present work has considered only the *static* behaviour of carriers in nanostructures. In other words, the states that carriers *could* occupy within a structure have been calculated, along with the Fermi–Dirac distributions of carriers in thermal equilibrium. However, the equilibrium distribution of carriers is broken in real heterostructure devices (e.g. by the injection of an electric current or an optical 'pump'), and it is therefore important to consider the *dynamic* behaviour of carriers in such systems in order to predict their distribution accurately.

In principle, a complete model of the behaviour of a carrier within a heterostructure *could* be obtained by finding a Hamiltonian that describes all possible interactions between the carrier and its surroundings, and then solving the time-dependent Schrödinger equation to observe the evolution of the carrier wave function over time. However, this is a formidable computational task, which is further complicated by the random nature of many of the interactions (e.g. randomly distributed impurities). To simplify the problem, carrier dynamics may be approximated by two distinct types of process: coherent processes in which wave functions evolve smoothly between their initial and final states; and incoherent, scattering processes. The latter are the focus of this chapter, and are defined as random, infrequent but very rapid events that transfer a carrier between two states. This could be caused by interactions between carriers and any kind of perturbing potential, including other carriers, impurities, and defects in the heterostructure, or with lattice vibrations (phonons) and photons.

Since scattering is assumed to be infrequent, the system can be assumed to stabilise to its static, unperturbed state between events. As such, the eigenstates of the unperturbed system can be calculated (as in Chapters 2 and 3) independently from the scattering dynamics. The very rapid nature of scattering events introduces a further simplification: carriers effectively 'hop' instantaneously from one state to another, and the only necessary task is to compute the *rate* of scattering (i.e. the number of events per second) for any combination of initial and final states, which is the focus of this chapter. After determining these scattering rates for appropriate interaction processes, the results are used in Chapter 12 to compute the populations of states and current density in devices.

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

10.2 Fermi's golden rule

As stated above, scattering is caused, in the most general sense, by an interaction between a carrier and a perturbing potential. In this case, the Hamiltonian may be written as:

$$\mathcal{H} = \mathcal{H}_0 + \tilde{\mathcal{H}} \tag{10.1}$$

where $\mathcal{H}_0$ is the Hamiltonian for the unperturbed system (i.e. without scattering), and $\tilde{\mathcal{H}}$ is an additional perturbing term that causes scattering to occur. The states of the unperturbed system can be computed using techniques from previous chapters, to find the energies E_j and the corresponding wave functions:

$$|j(t)\rangle = |j\rangle \exp\left(-\frac{iE_j^{\mathrm{t}}\, t}{\hbar}\right) \tag{10.2}$$

Note that the time-dependence of the wave function has been made explicit here, since the aim is to consider how the state *changes over time*. A superscript 't' on the energies has also been introduced here to indicate that these are the total carrier energies, which contain both kinetic and potential energy components.

The solution for the *perturbed* system is now obtained by solving the *time-dependent* Schrödinger equation (equation (1.20)):

$$[\mathcal{H}_0 + \tilde{\mathcal{H}}]\Psi(\mathbf{r}, t) = i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{r}, t) \tag{10.3}$$

The solutions of the time-dependent Schrödinger equation can be expressed at any given time as a series expansion using the solutions of the unperturbed system:

$$\Psi(\mathbf{r}, t) = \sum_j a_j(t)|j(t)\rangle, \tag{10.4}$$

where $a_j(t)$ are time-dependent weighting coefficients. Substituting this solution into equation (10.3) yields:

$$\sum_j [a_j(t)\mathcal{H}_0|j(t)\rangle + a_j(t)\tilde{\mathcal{H}}|j(t)\rangle] = i\hbar\sum_j \frac{\partial}{\partial t}\left[a_j(t)|j(t)\rangle\right] \tag{10.5}$$

The derivative on the right-hand side can be expanded using the product rule to give:

$$\sum_j [a_j(t)\mathcal{H}_0|j(t)\rangle + a_j(t)\tilde{\mathcal{H}}|j(t)\rangle] = i\hbar\sum_j \left[a_j(t)\frac{\partial}{\partial t}|j(t)\rangle + \frac{\mathrm{d}a_j(t)}{\mathrm{d}t}|j(t)\rangle\right] \tag{10.6}$$

Now, the first terms on each side are merely the solutions of the unperturbed system, and therefore cancel out, leaving:

$$\sum_j a_j(t)\tilde{\mathcal{H}}|j(t)\rangle = i\hbar\sum_j \left[\frac{\mathrm{d}a_j(t)}{\mathrm{d}t}|j(t)\rangle\right] \tag{10.7}$$

To compute the transition to a final state, $|f\rangle$, multiply both sides by the complex conjugate and integrate over all space to yield:

$$\sum_j a_j(t)\langle f(t)|\tilde{\mathcal{H}}|j(t)\rangle = i\hbar\sum_j\left[\frac{\mathrm{d}a_j(t)}{\mathrm{d}t}\langle f(t)|j(t)\rangle\right] \tag{10.8}$$

Now, noting the orthonormality of the states on the right-hand-side and substituting in the explicit time-dependence from equation (10.2) yields:

$$\sum_j a_j(t)\langle f|\tilde{\mathcal{H}}|j\rangle\exp\left[i\frac{(E_f^{\mathrm{t}}-E_j^{\mathrm{t}})}{\hbar}\right] = i\hbar\frac{\mathrm{d}a_f(t)}{\mathrm{d}t} \tag{10.9}$$

A further simplification can be made by assuming that if the perturbation is weak, and the carrier is initially located in state $|i\rangle$, then only the ith term in the summation will be significant, giving:

$$\langle f|\tilde{\mathcal{H}}|i\rangle\exp\left[i\frac{(E_f^{\mathrm{t}}-E_i^{\mathrm{t}})}{\hbar}\right] = i\hbar\frac{\mathrm{d}a_f(t)}{\mathrm{d}t} \tag{10.10}$$

Thus, an expression has been derived for the change in weighting of the final state with respect to time, in terms of a *scattering matrix element*, $\langle f|\tilde{\mathcal{H}}|i\rangle$, which describes how strongly a given pair of states is coupled by a given perturbation.

Finally, the *probability* of locating the carrier in the final state at a given time is found by integrating both sides of the equation and taking the squared magnitude:

$$|a_f(t)|^2 = \frac{1}{\hbar^2}\left|\int_0^t\langle f|\tilde{\mathcal{H}}|i\rangle\exp\left[\frac{i(E_f^{\mathrm{t}}-E_i^{\mathrm{t}})t'}{\hbar}\right]\,\mathrm{d}t'\right|^2 \tag{10.11}$$

The integral is clearly dependent on the scattering matrix element, and therefore cannot be simplified without further assumptions being made. However, if the scattering process is time-independent, then the expression can be simplified to:

$$|a_f(t)|^2 = \frac{1}{\hbar^2}|\langle f|\tilde{\mathcal{H}}|i\rangle|^2\left|\int_0^t\exp\left[\frac{i(E_f^{\mathrm{t}}-E_i^{\mathrm{t}})t'}{\hbar}\right]\,\mathrm{d}t'\right|^2 \tag{10.12}$$

The integral can be now solved analytically to obtain:

$$|a_f(t)|^2 = \frac{|\langle f|\tilde{\mathcal{H}}|i\rangle|^2t^2}{\hbar^2}\,\mathrm{sinc}^2\left(\frac{(E_f^{\mathrm{t}}-E_i^{\mathrm{t}})t}{2\hbar}\right). \tag{10.13}$$

Now, since the scattering event is assumed to be very rapid, it is reasonable to ignore the initial transients. The steady-state solution is then given by [1]:

$$\lim_{t\to\infty}|a_f(t)|^2 = \frac{2\pi|\langle f|\tilde{\mathcal{H}}|i\rangle|^2t}{\hbar}\delta\left(E_f^{\mathrm{t}}-E_i^{\mathrm{t}}\right). \tag{10.14}$$

The scattering rate (i.e. the rate at which carriers transfer from initial state $|i\rangle$ to final state $|f\rangle$) is then given by the derivative of this expression:

$$W_{if} = \frac{2\pi}{\hbar}\left|\langle f|\tilde{\mathcal{H}}|i\rangle\right|^2\delta(E_f^{\mathrm{t}}-E_i^{\mathrm{t}}). \tag{10.15}$$

Finally, the lifetime of the carrier in state $|i\rangle$ is given by summing over all the possible final states:

$$\frac{1}{\tau_i} = \frac{2\pi}{\hbar} \sum_f \left| \langle f|\tilde{\mathcal{H}}|i\rangle \right|^2 \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}}) \tag{10.16}$$

This important result is known as *Fermi's golden rule* [2] which relates a (static) perturbing potential to the lifetime of a carrier. It is important to note the Dirac delta function in this expression, which implies that carriers may only scatter to state $|f\rangle$ if that state lies at the same energy as the initial state. This demonstrates that static perturbations give rise to *elastic* scattering processes (i.e. where no energy is gained or lost by carriers).

10.3 Extension to sinusoidal perturbations

In addition to static perturbations, carriers can interact with 'wave-like' perturbations such as light (photons) or lattice vibrations (phonons). In this case, the scattering matrix element will take the form:

$$\langle f|\tilde{\mathcal{H}}(t)|i\rangle = \langle f|\tilde{\mathcal{H}}|i\rangle \left(\mathrm{e}^{+i\omega t} + \mathrm{e}^{-i\omega t}\right) \tag{10.17}$$

where the time-dependence has been separated from the magnitude of the perturbation, and ω is the angular frequency of the interacting wave.

The solution of equation (10.11) becomes [1]:

$$\begin{aligned} |a_f(t)|^2 = \frac{|\langle f|\tilde{\mathcal{H}}|i\rangle|^2 t^2}{\hbar^2} \Bigg[& \operatorname{sinc}^2 \frac{(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} - \hbar\omega)t}{2\hbar} + \operatorname{sinc}^2 \frac{(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} + \hbar\omega)t}{2\hbar} \\ & + 2\cos\omega t \operatorname{sinc} \frac{(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} - \hbar\omega)t}{2\hbar} \operatorname{sinc} \frac{(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} + \hbar\omega)t}{2\hbar} \Bigg]. \end{aligned} \tag{10.18}$$

The transition rate from the initial state $|i\rangle$ to the final state $|f\rangle$ is given by taking the steady-state limit ($t \to \infty$) [3]:

$$\frac{1}{\tau_i} = \frac{2\pi}{\hbar} \sum_f \left| \langle f|\tilde{\mathcal{H}}|i\rangle \right|^2 \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega) \tag{10.19}$$

The δ-function now explicitly contains the term $\hbar\omega$, and in general this represents the transfer of energy between a carrier and a boson (e.g. a photon or a phonon), with the upper sign of the $\mp$ representing absorption and the lower emission of a boson by the carrier. For example, for absorption the total initial energy is the sum of the carrier energy and the boson energy, i.e.

$$E_i^{\mathrm{t}} + \hbar\omega = E_f^{\mathrm{t}}, \qquad \text{and hence} \qquad E_f^{\mathrm{t}} - E_i^{\mathrm{t}} - \hbar\omega = 0 \tag{10.20}$$

and for emission:

$$E_i^{\mathrm{t}} = E_f^{\mathrm{t}} + \hbar\omega, \qquad \text{and hence} \qquad E_f^{\mathrm{t}} - E_i^{\mathrm{t}} + \hbar\omega = 0 \tag{10.21}$$

10.4 Averaging over two-dimensional carrier distributions

Equations (10.16) and (10.19) give general expressions for the lifetime of a single carrier scattering from an initial state i to all possible final states f. However, in real situations, there are many carriers distributed within energy bands (or subbands) within a system, and it is often more useful to know the *mean* scattering rate (or lifetime) for carriers scattering between a given pair of bands.

For a two-dimensional system (i.e. the focus of the present work), the scattering rate for a single carrier can be denoted $W_{if}(k_i)$, which indicates that the carrier is scattered from a particular state with wave vector $\mathbf{k}_i$ in the initial subband i to *all* possible states in the final subband f. Fermi's golden rule can then be written in the alternative form:

$$W_{if}(k_i) = \frac{2\pi}{\hbar} \sum_{\mathbf{k}_f} \left| \langle f | \tilde{\mathcal{H}}(\mathbf{k}_i, \mathbf{k}_f) | i \rangle \right|^2 \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}}) \tag{10.22}$$

for static perturbations, or

$$W_{if}(k_i) = \frac{2\pi}{\hbar} \sum_{\mathbf{k}_f} \left| \langle f | \tilde{\mathcal{H}}(\mathbf{k}_i, \mathbf{k}_f) | i \rangle \right|^2 \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega) \tag{10.23}$$

for sinusoidal potentials. These forms are identical to equations (10.16) and (10.19) respectively, except the set of destination states is restricted to those in subband f, and the wave-vector dependence of the scattering matrix element has been made explicit. Note that the scattering rate depends only on the magnitude of the initial wave vector, since it is assumed that all perturbations are always isotropic in the x–y plane.

The intersubband scattering rates can be calculated for *any* scattering process in a two-dimensional system, as long as an appropriate final wave vector $\mathbf{k}_f$ and scattering matrix element can be found for a given initial wave vector. It is shown later in this chapter that carriers usually scatter much faster *within* a subband than they transfer between subbands. As such, it is reasonable to assume that the carriers settle rapidly into a local thermal equilibrium within each subband, which gives rise to a Fermi–Dirac distribution (see Section 2.5).

A simple weighted mean over a distribution of carriers in the initial subband might look like the following:

$$\overline{W_{if}} = \frac{\int W_{if}(k_i) f_i^{\mathrm{FD}}(E_i^k) \ \mathrm{d}E_i^k}{\int f_i^{\mathrm{FD}}(E_i^k) \ \mathrm{d}E_i^k} \tag{10.24}$$

where the superscript k has been introduced to indicate that this is the in-plane kinetic energy associated with the wave vector $\mathbf{k}_i$, and the subscript i on the distribution functions indicates the subband, i.e. to be evaluated with the 'quasi'-Fermi energy of that subband.

By using the result of equation (2.69) in Section 2.5, the denominator is then given by:

$$\int f_i^{\mathrm{FD}}(E_i^k) \ \mathrm{d}E_i^k = N_i \frac{\pi\hbar^2}{m^*} \tag{10.25}$$

As the carrier's in-plane energy is expressed in terms of a wave vector and:

$$E_i^k = \frac{\hbar^2 k_i^2}{2m^*}, \quad \text{then} \quad \mathrm{d}E_i^k = \frac{\hbar^2 k_i}{m^*} \ \mathrm{d}k_i \tag{10.26}$$

Hence the above mean can be expressed as:

$$\overline{W_{if}} = \frac{\int W_{if}(k_i) f_i^{\mathrm{FD}}(k_i)\ k_i\ \mathrm{d}k_i}{\pi N_i} \tag{10.27}$$

However, this still disregards the distribution in the final subband; in this case, filled states could prevent carriers from scattering into them, thus reducing the probability of an event. This effect of *final-state blocking* can be incorporated into the above to give:

$$\overline{W_{if}} = \frac{\int W_{if}(k_i) f_i^{\mathrm{FD}}(k_i)[1 - f_f^{\mathrm{FD}}(k_f)]\ k_i\ \mathrm{d}k_i}{\pi N_i} \tag{10.28}$$

The integral is, in principle, evaluated from zero wave vector up to infinity, although in practice the upper limit will be set to some defined maximum 'cut-off' wave vector. One possible choice of cut-off could be the wave vector that corresponds to the energy of the highest point in the potential profile, which means physically that any carriers above the barrier are assumed to ionise rapidly. An alternative to this may be to choose the wave vector that corresponds to the Fermi energy plus $10kT$, or similar.

10.5 Phonons

The massive atoms that constitute semiconductor crystals are all connected together by chemical bonds, which are nominally covalent, although in compounds can have a degree of ionicity. These atoms are always in a state of continual motion, which because of the definite crystal lattice structure, is vibrational around an equilibrium position. The atoms vibrate even at the hypothetical zero of absolute temperature—the so-called *zero-point energy* (see Section 3.10). In some ways, the vibrations of these interconnected quantum particles (atoms) resemble a classical (macroscopic) system of a series of masses connected by springs. There are basically four different modes of vibration, as illustrated in Fig. 10.1, each one of which is referred to as a *phonon*.

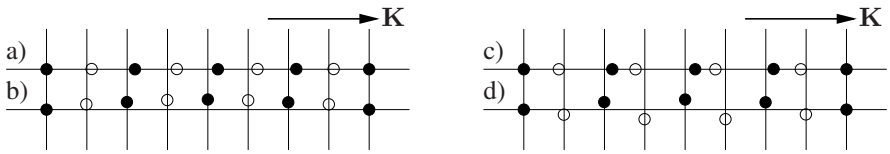


Figure 10.1: Schematic illustrations of the atomic displacements in (a) longitudinal acoustic (LA), (b) transverse acoustic (TA), (c) longitudinal optic (LO) and (d) transverse optic (TO) phonon modes

The acoustic modes shown in Fig. 10.1(a) and (b) are characterised by the neighbouring atoms being in phase. In the longitudinal mode the atomic displacements are in the same direction as the direction of energy transfer, while in the transverse mode the atomic displacements are perpendicular to this direction.

The longitudinal and transverse definitions also apply to the two types of optic phonon mode as illustrated in Fig. 10.1(c) and (d). However, in this type of lattice vibration the displacements of neighbouring atoms are in opposite phase.

The wave-like nature of the lattice vibrations allows them to be described, say, by an angular frequency ω and a wave vector $\mathbf{K}$. Thus the energy of a phonon is $\hbar\omega$—the same as a photon of light. In addition, and in analogy to propagating electrons, the momentum of a phonon is said to be quantised and of value $\hbar\mathbf{K}$. Furthermore, phonons are diffracted by the crystal lattice just like electrons and holes, and thus a Brillouin zone type summary of the energy–momentum curves can be employed.

Figure 10.2 shows schematically just such a set of phonon dispersion curves for a typical semiconductor. The forms of the curves are reasonably similar, although with differing energy scales for the common semiconductors that are of interest in this work (see, for example, [4], p. 14); note, however, that the coupling (interaction) between the different phonon modes with charge carriers does differ between the various materials.

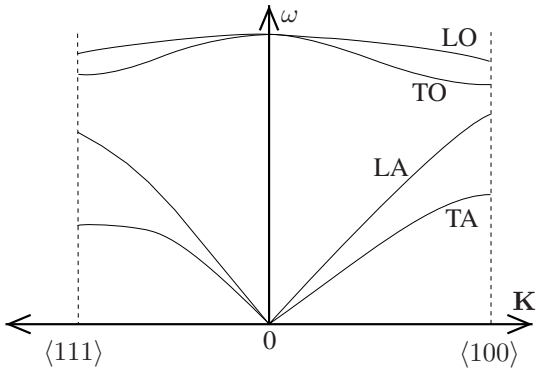


Figure 10.2: Phonon energy ($\hbar\omega$) versus momentum ($\hbar K$) curves for a typical semiconductor

Phonons are bosons, and hence their number per unit volume is given by the Bose–Einstein factor (see [5], p. 391 or [6], p. 454):

$$N_0 = \frac{1}{\exp\left(\hbar\omega/kT\right) - 1} \tag{10.29}$$

As the phonons themselves represent the motion of atoms that are centres of electric charge, they also represent time-dependent perturbations $\tilde{\mathcal{H}}$ of the crystal potential and can therefore scatter charge carriers. Many authors have previously considered scattering via phonons (see, for example, [7, 8, 9]); however, the spirit of this present work is to provide a *fully* documented derivation.

With this aim, consider the oscillating potential associated with a phonon in bulk material:[1]

$$\phi = C(\mathbf{K})\mathrm{e}^{-i\mathbf{K}\bullet\mathbf{r}} \tag{10.30}$$

[1]The authors would like to thank Paul Kinsler, University of Leeds, for his contributions to the derivation of phonon-scattering rates here and throughout the following sections.

Therefore, the electric field, which is the derivative of this oscillating potential, can be described by the following relationship:

$$\mathbf{E} = \nabla\phi = -i\mathbf{K}\phi \tag{10.31}$$

and hence:

$$\mathbf{E}^*\mathbf{E} = |C(\mathbf{K})|^2\,\mathbf{K}_\bullet\mathbf{K} \tag{10.32}$$

The normalisation condition is therefore:

$$\frac{1}{2}\hbar\omega(\mathbf{K}) = \frac{1}{2}\omega(\mathbf{K})\left.\frac{\partial\epsilon}{\partial\omega}\right|_{\mathbf{K}}\int \mathbf{E}^*\mathbf{E}\,\mathrm{d}\tau \tag{10.33}$$

If Ω is the volume of the crystal, then as the integrand is independent of position (see equation (10.32)):

$$\frac{1}{2}\hbar\omega(\mathbf{K}) = \frac{1}{2}\omega(\mathbf{K})\left.\frac{\partial\epsilon}{\partial\omega}\right|_{\mathbf{K}}\Omega\,|C(\mathbf{K})|^2\,\mathbf{K}_\bullet\mathbf{K} \tag{10.34}$$

The majority of the interest here lies with heterostructures made from compound semiconductors. These materials are polar as the different electronegativities of the constituent atoms lead to a degree of ionicity in the chemical bonds (see Chapter 15). In such materials the dominant electron–phonon interaction (scattering) is with the *longitudinal optic phonon*, often referred to as the LO phonon. The LO phonon dispersion curve (see Fig. 10.2) is relatively flat, and hence it is possible to approximate it as being *dispersionless*, which gives $\omega\partial\epsilon/\partial\omega = 2/P$, where:

$$P = \frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_\mathrm{s}} \tag{10.35}$$

with ϵ_∞ and ϵ_s being the high- and low-frequency permittivities of the material, respectively. Therefore:

$$\frac{1}{2}\hbar\omega(\mathbf{K}) = \frac{1}{2}\frac{2}{P}\Omega\,|C(\mathbf{K})|^2\,\mathbf{K}_\bullet\mathbf{K} \tag{10.36}$$

The normalisation coefficients $C(\mathbf{K})$ of the phonon potential wave are therefore given by:

$$|C(\mathbf{K})|^2 = \frac{\hbar\omega P}{2\Omega|\mathbf{K}|^2} \tag{10.37}$$

i.e. the normalised potential wave associated with a single dispersionless phonon is:

$$\phi = \left(\frac{\hbar\omega P}{2|\mathbf{K}|^2}\right)^{\frac{1}{2}}\frac{\mathrm{e}^{-i\mathbf{K}_\bullet\mathbf{r}}}{\Omega^{\frac{1}{2}}} \tag{10.38}$$

The total *phonon interaction term* is thus obtained by summing over all phonon wave vectors, i.e.

$$\tilde{\mathcal{H}} = e\sum_{\mathbf{K}}\phi \tag{10.39}$$

and therefore:

$$\tilde{\mathcal{H}} = e\sum_{\mathbf{K}}\left(\frac{\hbar\omega P}{2|\mathbf{K}|^2}\right)^{\frac{1}{2}}\frac{\mathrm{e}^{-i\mathbf{K}_\bullet\mathbf{r}}}{\Omega^{\frac{1}{2}}} \tag{10.40}$$

10.6 Longitudinal optic phonon scattering of two-dimensional carriers

The phonon interaction term is as for the case of bulk phonon modes, as defined in equation (10.40). However, in order to make use of the symmetry of a general heterostructure, it can be split into components along the growth (z-)axis and in the x–y plane of the layers, i.e.

$$\tilde{\mathcal{H}} = e \sum_{\mathbf{K}_{xy}} \sum_{K_z} \left(\frac{\hbar\omega P}{2\left(|\mathbf{K}_{xy}|^2 + |K_z|^2\right)} \right)^{\frac{1}{2}} \frac{\mathrm{e}^{-i\mathbf{K}_{xy}\bullet\mathbf{r}_{xy}}}{A^{\frac{1}{2}}} \frac{\mathrm{e}^{-iK_z z}}{L^{\frac{1}{2}}} \tag{10.41}$$

Now the electron (or hole) wave functions in a heterostructure are a product of an envelope along the growth axis and an in-plane travelling wave (see Section 2.2), i.e.

$$\psi = \psi(z) \frac{\mathrm{e}^{-i\mathbf{k}_\bullet \mathbf{r}_{xy}}}{A^{\frac{1}{2}}} \tag{10.42}$$

Therefore, substituting this new form for the phonon interaction and the electron wave function into Fermi's golden rule (equation 10.23) gives:

$$W_{if}(k_i) = \frac{2\pi}{\hbar} \sum_{\mathbf{k}_f} \left| \iint \psi_f^*(z) \frac{\mathrm{e}^{i\mathbf{k}_f\bullet\mathbf{r}_{xy}}}{A^{\frac{1}{2}}} e \sum_{\mathbf{K}_{xy}} \sum_{K_z} \left(\frac{\hbar\omega P}{2\left(|\mathbf{K}_{xy}|^2 + |K_z|^2\right)} \right)^{\frac{1}{2}} \right.$$

$$\left. \times \frac{\mathrm{e}^{-i\mathbf{K}_{xy}\bullet\mathbf{r}_{xy}}}{A^{\frac{1}{2}}} \frac{\mathrm{e}^{-iK_z z}}{L^{\frac{1}{2}}} \psi_i(z) \frac{\mathrm{e}^{-i\mathbf{k}_i\bullet\mathbf{r}_{xy}}}{A^{\frac{1}{2}}} \, \mathrm{d}z \, \mathrm{d}\mathbf{r}_{xy} \right|^2 \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega) \tag{10.43}$$

where the electron wave vectors, $\mathbf{k}_i$ and $\mathbf{k}_f$, of the initial and final states respectively, are taken explicitly to lie in the plane of the quantum wells only, i.e. $\mathbf{k} = \mathbf{k}(x, y)$ only.

The above represents the lifetime for scattering from a total phonon population of 1, and although the sum over all of the phonon wave vectors should introduce a population term, it is simpler just to add it here manually. This is done with the factor $(N_0 + \frac{1}{2} \mp \frac{1}{2})$ which represents the phonon density within the crystal. The upper sign of the $\mp$ represents absorption, which reduces the phonon population from $N_0 + 1$ to N_0, while the lower sign represents emission of a phonon which increases the number of phonons from N_0 to $N_0 + 1$. Just for convenience, absorb this factor into P and rewrite as P', i.e.

$$P' = \left(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_{\mathrm{s}}} \right) \left(N_0 + \frac{1}{2} \mp \frac{1}{2} \right) \tag{10.44}$$

This then gives:

$$W_{if}(k_i) = \frac{2\pi e^2 \hbar\omega P'}{2\hbar A} \sum_{\mathbf{k}_f} \left| \sum_{K_z} \frac{1}{L^{\frac{1}{2}}} \int \psi_f^*(z) \mathrm{e}^{-iK_z z} \psi_i(z) \, \mathrm{d}z \right.$$

$$\left. \times \sum_{\mathbf{K}_{xy}} \frac{1}{\left(|\mathbf{K}_{xy}|^2 + |K_z|^2\right)^{\frac{1}{2}}} \frac{1}{A} \int \mathrm{e}^{-i(\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}_{xy})\bullet\mathbf{r}_{xy}} \, \mathrm{d}\mathbf{r}_{xy} \right|^2 \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega) \tag{10.45}$$

The integral over the x–y plane (denoted by $\mathrm{d}\mathbf{r}_{xy}$) of the exponential function $\exp\left[-i(\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}_{xy})_\bullet \mathbf{r}_{xy}\right]$ can be converted into a δ-function, provided that the limits of integration are effectively at infinity. This gives a factor of 2π per dimension, resulting in:

$$W_{if}(k_i) = \frac{\pi e^2 \omega P'}{AL} \sum_{\mathbf{k}_f} \left| \sum_{K_z} G_{if}(K_z) \right.$$

$$\left. \times \sum_{\mathbf{K}_{xy}} \frac{1}{(|\mathbf{K}_{xy}|^2 + |K_z|^2)^{\frac{1}{2}}} \frac{(2\pi)^2}{A} \delta(\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}_{xy}) \right|^2 \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega) \tag{10.46}$$

where $G_{if}(K_z)$ is known as the *form factor* and is given by:

$$G_{if}(K_z) = \int \psi_f^*(z) \mathrm{e}^{-iK_z z} \psi_i(z)\, \mathrm{d}z \tag{10.47}$$

Often this form factor is normalised by dividing by some length, or square root of a length (see, for example, Lundstrom [8], equation (2.102)). However, it suits the purpose better here to leave it unnormalised as above.

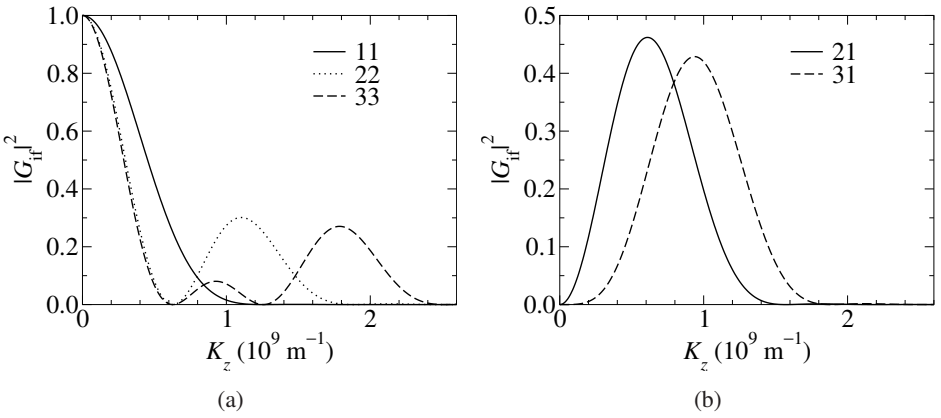


Figure 10.3: Form factors for (a) intrasubband and (b) intersubband phonon scattering

Figure 10.3(a) shows this form factor for the case of *intrasubband* scattering events, i.e. transitions where the carrier remains within the same quantum well subband. For this particular case, the example system is that of an electron in a 100 Å GaAs infinitely deep quantum well. As the form factor $G_{if}(K_z)$ always appears as a modulus squared, i.e. $|G_{if}(K_z)|^2$ (see later), and as the initial- and final-state wave functions, ψ_i and ψ_f, in these confined systems are real, $|G_{if}(K_z)|^2$ is symmetric about $K_z = 0$ and hence is plotted for positive K_z only.

The figure shows that $|G_{if}(K_z)|^2$ can have more than one local maximum, with the overall maximum always for the zone-centre ($K_z = 0$) phonon, which implies that this scattering

event is most likely to occur. The physical interpretation is that as the initial and final states of the carrier are the same, then no momentum change along the growth (z-)axis is required, and hence the dominance of the zero-momentum transition. There are additional local maxima, the number of which depends on the number of antinodes in the wave functions. The form factor squared for the $2 \rightarrow 2$ event, for example, has a total of two maxima, i.e. one at the zone centre and another local maximum. This second local maximum *may* be thought of as corresponding to a scattering event from one of the wave function nodes to the other.

Figure 10.3(b) shows the corresponding case for *intersubband* transitions, i.e. scattering events where the carrier changes from one confined level to another. These form factors are quite different in their behaviour than those of the intrasubband case in that there is no maximum at the zone centre—rather it is at a finite wave vector. The particular value corresponds to the difference in momenta along the z-axis of the carrier states. Thus, as i increases, the wave vector of the antinode of the $i \rightarrow 1$ transition also increases.

Converting the summations over the phonon wave vectors $\mathbf{K}_{xy}$ and K_z into integrals introduces factors of $A/(2\pi)^2$ and $L/(2\pi)$, respectively, from the density of states, and therefore:

$$W_{if}(k_i) = \frac{\pi e^2 \omega P'}{AL} \sum_{\mathbf{k}_f} \left| \frac{L}{2\pi} \iint G_{if}(K_z) \frac{\delta(\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}_{xy})}{(|\mathbf{K}_{xy}|^2 + |K_z|^2)^{\frac{1}{2}}} \, \mathrm{d}K_z \, \mathrm{d}\mathbf{K}_{xy} \right|^2 \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega) \tag{10.48}$$

Finally, changing the summation over the final in-plane electron wave vector $\mathbf{k}_f$ into an integral, and also introducing a factor of $A/(2\pi)^2$, yields:

$$W_{if}(k_i) = \frac{\pi e^2 \omega P' L}{A(2\pi)^2} \frac{A}{(2\pi)^2} \int \left| \iint G_{if}(K_z) \frac{\delta(\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}_{xy})}{(|\mathbf{K}_{xy}|^2 + |K_z|^2)^{\frac{1}{2}}} \, \mathrm{d}K_z \, \mathrm{d}\mathbf{K}_{xy} \right|^2 \mathrm{d}\mathbf{k}_f \times \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega) \tag{10.49}$$

Expanding the modulus squared doubles the number of integrals over the phonon wave vectors, i.e.

$$W_{if}(k_i) = \frac{\pi e^2 \omega P' L}{(2\pi)^4} \int \left(\iint G_{if}^*(K_z) \frac{\delta(\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}_{xy})}{(|\mathbf{K}_{xy}|^2 + |K_z|^2)^{\frac{1}{2}}} \, \mathrm{d}K_z \, \mathrm{d}\mathbf{K}_{xy} \right) \tag{10.50}$$

$$\times \left(\iint G_{if}(K_z') \frac{\delta(\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}_{xy}')}{\left(|\mathbf{K}_{xy}'|^2 + |K_z'|^2\right)^{\frac{1}{2}}} \, \mathrm{d}K_z' \, \mathrm{d}\mathbf{K}_{xy}' \right) \mathrm{d}\mathbf{k}_f \, \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega)$$

Change the order of the integrations in such a way as to perform the integration over $\mathbf{K}_{xy}'$ first, i.e.

$$W_{if}(k_i) = \frac{\pi e^2 \omega P' L}{(2\pi)^4} \iiint G_{if}^*(K_z) G_{if}(K_z') \iint \frac{\delta(\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}_{xy})}{(|\mathbf{K}_{xy}|^2 + |K_z|^2)^{\frac{1}{2}}}$$

$$\times \frac{\delta(\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}_{xy}')}{\left(|\mathbf{K}_{xy}'|^2 + |K_z'|^2\right)^{\frac{1}{2}}} \, \mathrm{d}\mathbf{K}_{xy}' \, \mathrm{d}\mathbf{k}_f \, \mathrm{d}K_z' \, \mathrm{d}K_z \, \mathrm{d}\mathbf{K}_{xy} \, \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega) \tag{10.51}$$

So, by performing this integration over $\mathbf{K}'_{xy}$, the second δ-function limits the contribution to the point where $\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}'_{xy} = 0$, i.e. when $\mathbf{K}'_{xy} = -\mathbf{k}_i + \mathbf{k}_f$, thus giving:

$$W_{if}(k_i) = \frac{\pi e^2 \omega P' L}{(2\pi)^4} \iiint G^*_{if}(K_z) G_{if}(K'_z) \int \frac{\delta(\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}_{xy})}{(|\mathbf{K}_{xy}|^2 + |K_z|^2)^{\frac{1}{2}}} \tag{10.52}$$

$$\times \frac{1}{(|-\mathbf{k}_i + \mathbf{k}_f|^2 + |K'_z|^2)^{\frac{1}{2}}} \, \mathrm{d}\mathbf{k}_f \, \mathrm{d}K'_z \, \mathrm{d}K_z \, \mathrm{d}\mathbf{K}_{xy} \, \delta(E^{\mathrm{t}}_f - E^{\mathrm{t}}_i \mp \hbar\omega)$$

Now when performing the integration over the final electron in-plane wave vector $\mathbf{k}_f$, as denoted by the $\mathrm{d}\mathbf{k}_f$, again its contribution is limited by the remaining δ-function to the point where $\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}_{xy} = 0$, i.e. when $\mathbf{k}_f = \mathbf{k}_i + \mathbf{K}_{xy}$, thus giving:

$$W_{if}(k_i) = \frac{\pi e^2 \omega P' L}{(2\pi)^4} \iiint G^*_{if}(K_z) G_{if}(K'_z) \frac{1}{(|\mathbf{K}_{xy}|^2 + |K_z|^2)^{\frac{1}{2}}}$$

$$\times \frac{1}{(|-\mathbf{k}_i + \mathbf{k}_i + \mathbf{K}_{xy}|^2 + |K'_z|^2)^{\frac{1}{2}}} \, \mathrm{d}K'_z \, \mathrm{d}K_z \, \mathrm{d}\mathbf{K}_{xy} \, \delta(E^{\mathrm{t}}_f - E^{\mathrm{t}}_i \mp \hbar\omega) \tag{10.53}$$

and therefore:

$$W_{if}(k_i) = \frac{\pi e^2 \omega P' L}{(2\pi)^4} \iiint \frac{G^*_{if}(K_z) \, G_{if}(K'_z)}{(|\mathbf{K}_{xy}|^2 + |K_z|^2)^{\frac{1}{2}} \, (|\mathbf{K}_{xy}|^2 + |K'_z|^2)^{\frac{1}{2}}} \, \mathrm{d}K'_z \, \mathrm{d}K_z \, \mathrm{d}\mathbf{K}_{xy}$$

$$\times \; \delta(E^{\mathrm{t}}_f - E^{\mathrm{t}}_i \mp \hbar\omega) \tag{10.54}$$

It is now necessary to perform a little mathematical trickery in order to force the problem through and obtain the *accepted* result (see, for example, Lundstrom [8]). Start by rearranging equation (10.54) to give:

$$W_{if}(k_i) = \frac{\pi e^2 \omega P' L}{(2\pi)^4} \iint \frac{G^*_{if}(K_z)}{(|\mathbf{K}_{xy}|^2 + |K_z|^2)^{\frac{1}{2}}} \, \mathrm{d}K_z \int \frac{G_{if}(K'_z)}{(|\mathbf{K}_{xy}|^2 + |K'_z|^2)^{\frac{1}{2}}} \, \mathrm{d}K'_z$$

$$\times \; \mathrm{d}\mathbf{K}_{xy} \, \delta(E^{\mathrm{t}}_f - E^{\mathrm{t}}_i \mp \hbar\omega) \tag{10.55}$$

Then:

$$W_{if}(k_i) = \frac{\pi e^2 \omega P' L}{(2\pi)^4} \iint \frac{G^*_{if}(K_z)}{(|\mathbf{K}_{xy}|^2 + |K_z|^2)^{\frac{1}{2}}}$$

$$\times \left[\int \frac{G_{if}(K_z)}{(|\mathbf{K}_{xy}|^2 + |K_z|^2)^{\frac{1}{2}}} \, \frac{2\pi}{L} \delta(K_z - K'_z) \right] \mathrm{d}K_z \, \mathrm{d}K'_z \, \mathrm{d}\mathbf{K}_{xy} \, \delta(E^{\mathrm{t}}_f - E^{\mathrm{t}}_i \mp \hbar\omega) \tag{10.56}$$

Note that the δ-function ensures that the term in the square brackets only gives a contribution when $K_z = K'_z$, and, when it does, this term reverts to the integral over K'_z, as before. An extra factor of $2\pi/L$ has been introduced here to satisfy dimensionality arguments, since the

δ-function has the dimensions of length. Changing the order of integration and simplifying the constant factors:

$$W_{if}(k_i) = \frac{\pi e^2 \omega P'}{(2\pi)^3} \iint \frac{G_{if}^*(K_z) G_{if}(K_z)}{(|\mathbf{K}_{xy}|^2 + |K_z|^2)^{\frac{1}{2}} \, (|\mathbf{K}_{xy}|^2 + |K_z|^2)^{\frac{1}{2}}} \times \int \delta(K_z - K_z') \, \mathrm{d}K_z' \, \mathrm{d}K_z \, \mathrm{d}\mathbf{K}_{xy} \, \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega) \tag{10.57}$$

Performing the integration over K_z' first,the δ-function has to yield a constant term. Furthermore, introducting the notation $|\mathbf{K}| = K$ gives:

$$W_{if}(k_i) = \frac{\pi e^2 \omega P'}{(2\pi)^3} \iint \frac{|G_{if}(K_z)|^2}{K_{xy}^2 + K_z^2} \, \mathrm{d}K_z \, \mathrm{d}\mathbf{K}_{xy} \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega) \tag{10.58}$$

The conservation of energy in the remaining δ-function must now be considered. The $\mp\hbar\omega$ term represents the energy of the phonon, with the upper sign corresponding to absorption of a phonon, and the lower sign representing emission. The total electron (or hole) energies are a sum of a subband minimum (labelled E_f and E_i) and the kinetic energy within the subband:

$$E_i^{\mathrm{t}} = E_i + \frac{\hbar^2 k_i^2}{2m^*} \tag{10.59}$$

and

$$E_f^{\mathrm{t}} = E_f + \frac{\hbar^2 k_f^2}{2m^*} \tag{10.60}$$

The optical branches of the phonon dispersion curves in typical III–V materials are quite flat (see Fig. 10.2), which means that the energy $\hbar\omega$ of the LO phonon is only a weak function of the phonon wave vector $\mathbf{K}$ (see, for example, [6], p. 435 and [10], p. 70). Hence, this derivation will follow the standard assumption that the phonon energy $\hbar\omega$ can be approximated well with a constant value (taken as $36\,\mathrm{meV}$ in GaAs ([10], p. 92)). It must be noted that the assumption of parabolic subbands made in the previous equations could limit the range of applicability for the case of holes in the valence band. Exploiting the conservation of energy, as specified explicitly in the δ-function, by substituting E_f^{t} and E_i^{t} into equation (10.58) then gives:

$$W_{if}(k_i) = \frac{\pi e^2 \omega P'}{(2\pi)^3} \iint \frac{|G_{if}(K_z)|^2}{K_{xy}^2 + K_z^2} \delta\left(E_f + \frac{\hbar^2 k_f^2}{2m^*} - E_i - \frac{\hbar^2 k_i^2}{2m^*} \mp \hbar\omega\right) \mathrm{d}K_z \, \mathrm{d}\mathbf{K}_{xy} \tag{10.61}$$

Labelling $E_f - E_i \mp \hbar\omega$ as Δ (say), then:

$$W_{if}(k_i) = \frac{\pi e^2 \omega P'}{(2\pi)^3} \iint \frac{|G_{if}(K_z)|^2}{K_{xy}^2 + K_z^2} \delta\left(\frac{\hbar^2 k_f^2}{2m^*} - \frac{\hbar^2 k_i^2}{2m^*} + \Delta\right) \mathrm{d}K_z \, \mathrm{d}\mathbf{K}_{xy} \tag{10.62}$$

The physical requirement of conservation of energy has therefore advanced the mathematical derivation, and now it is the turn of the conservation of momentum. Consider the application of the cosine rule to Fig. 10.4, remembering that as the initial and final

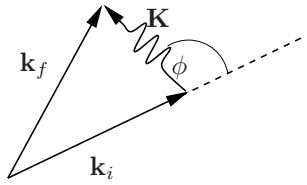


Figure 10.4: Momentum conservation in phonon scattering

momenta of the carrier are in-plane, the phonon momentum **K** must also be in the plane. The latter feature has been specified up until now with the additional index xy; thus:

$$k_f^2 = k_i^2 + K_{xy}^2 + 2k_i K_{xy}\cos\phi \tag{10.63}$$

Substituting for k_f^2 in the δ-function therefore gives:

$$W_{if}(k_i) = \frac{\pi e^2 \omega P'}{(2\pi)^3} \iint \frac{|G_{if}(K_z)|^2}{K_{xy}^2 + K_z^2}$$
$$\times\ \delta\left(\frac{\hbar^2}{2m^*}\left(k_i^2 + K_{xy}^2 + 2k_i K_{xy}\cos\phi\right) - \frac{\hbar^2 k_i^2}{2m^*} + \Delta\right)\ \mathrm{d}K_z\ \mathrm{d}\mathbf{K}_{xy} \tag{10.64}$$

Now simplifying the argument of the δ-function:

$$W_{if}(k_i) = \frac{\pi e^2 \omega P'}{(2\pi)^3} \iint \frac{|G_{if}(K_z)|^2}{K_{xy}^2 + K_z^2} \delta\left(\frac{\hbar^2 K_{xy}^2}{2m^*} + \frac{\hbar^2 k_i K_{xy}\cos\phi}{m^*} + \Delta\right)\ \mathrm{d}K_z\ \mathrm{d}\mathbf{K}_{xy} \tag{10.65}$$

Using $\delta(ax) = \delta(x)/a$, then:

$$W_{if}(k_i) = \frac{\pi e^2 \omega P'}{(2\pi)^3} \iint \frac{|G_{if}(K_z)|^2}{K_{xy}^2 + K_z^2} \frac{2m^*}{\hbar^2} \delta\left(K_{xy}^2 + 2k_i K_{xy}\cos\phi + \frac{2m^*\Delta}{\hbar^2}\right)\ \mathrm{d}K_z\ \mathrm{d}\mathbf{K}_{xy} \tag{10.66}$$

Now the two-dimensional integral over the Cartesian in-plane phonon wave vector, as denoted by $\mathrm{d}\mathbf{K}_{xy}$, can be changed into polar coordinates, with the radius as the modulus K_{xy} and an angle ϕ between it and the initial carrier momentum state. The new elemental area is $K_{xy}\ \mathrm{d}K_{xy}\ \mathrm{d}\phi$, and therefore:

$$W_{if}(k_i) = \frac{m^* e^2 \omega P'}{(2\pi)^2 \hbar^2} \int_0^{2\pi} \int_0^{\infty} \int_{-\infty}^{+\infty} \frac{|G_{if}(K_z)|^2}{K_{xy}^2 + K_z^2}$$
$$\times\ \delta\left(K_{xy}^2 + 2k_i K_{xy}\cos\phi + \frac{2m^*\Delta}{\hbar^2}\right) K_{xy}\ \mathrm{d}K_z\ \mathrm{d}K_{xy}\ \mathrm{d}\phi \tag{10.67}$$

This analysis could be continued by introducing a new variable $y = \cos\phi$; however, in this work we will follow the method of Hagston, Piorek and Harrison (for more details of this and related work, see the PhD thesis of Piorek [11]). This will involve factorising the argument of the δ-function. For now, however, consider the integral over ϕ first.

As $\cos\phi$ is an even function, the integral from 0 to 2π is clearly twice the integral from 0 to π, and, in addition, changing the order of integration gives:

$$W_{if}(k_i) = \frac{2m^* e^2 \omega P'}{(2\pi)^2 \hbar^2} \int_{-\infty}^{+\infty} \int_0^{\infty} \int_0^{\pi} \frac{|G_{if}(K_z)|^2}{K_{xy}^2 + K_z^2} \times \delta\left(K_{xy}^2 + 2k_i K_{xy}\cos\phi + \frac{2m^*\Delta}{\hbar^2}\right) K_{xy}\,\mathrm{d}\phi\,\mathrm{d}K_{xy}\,\mathrm{d}K_z \tag{10.68}$$

Write the constant prefactor as:

$$\Upsilon'' = \frac{2m^* e^2 \omega P'}{(2\pi)^2 \hbar^2} \tag{10.69}$$

and, in addition, consider the case of a positive Δ. When $0 < \phi < \pi/2$, the factor $\cos\phi$ is positive, which implies that the argument of the δ-function can only be factorised into the form $(K_{xy} + \alpha_1)(K_{xy} + \alpha_2)$, where α_1 and α_2 are real and positive. As K_{xy} is the *magnitude* of the phonon wave vector in the plane of the quantum wells, it is always positive; thus, the argument of the δ-function is never zero and no contributions to the integral are made.

However, when $\pi/2 < \phi < \pi$, the factor $\cos\phi$ is negative. Consider the substitution $\phi' = \pi - \phi$, and then $\cos\phi$ becomes $-\cos\phi'$ and $\mathrm{d}\phi = -\mathrm{d}\phi'$. When $\phi = \pi/2$, $\phi' = \pi/2$ and when $\phi = \pi$, $\phi' = 0$. Therefore:

$$W_{if}(k_i) = \Upsilon'' \int_{-\infty}^{+\infty} \int_0^{\infty} \int_0^{\pi/2} \frac{|G_{if}(K_z)|^2}{K_{xy}^2 + K_z^2} \times \delta\left(K_{xy}^2 - 2k_i K_{xy}\cos\phi' + \frac{2m^*\Delta}{\hbar^2}\right) K_{xy}\,\mathrm{d}\phi'\,\mathrm{d}K_{xy}\,\mathrm{d}K_z \tag{10.70}$$

Now the argument of the quadratic can be factorised into the form $(K_{xy} - \alpha_1)(K_{xy} - \alpha_2)$, where it is also specified that $\alpha_1 > \alpha_2$. Thus two solutions for the in-plane phonon wave vector exist; this is illustrated for a general case in Fig. 10.5. For each absorption or emission process, two possible scattering events can occur, each of which conserves both energy and momentum, where the latter are satisfied by mediating via two different phonon momenta, K_{xy}.

Using this factorised form in equation (10.70) gives:

$$W_{if}(k_i) = \Upsilon'' \int_{-\infty}^{+\infty} \int_0^{\infty} \int_0^{\pi/2} \frac{|G_{if}(K_z)|^2}{K_{xy}^2 + K_z^2} \delta\left((K_{xy} - \alpha_1)(K_{xy} - \alpha_2)\right) K_{xy}\,\mathrm{d}\phi'\,\mathrm{d}K_{xy}\,\mathrm{d}K_z \tag{10.71}$$

Now around the solution $K_{xy} = \alpha_1$, the factor $K_{xy} - \alpha_2$ is a constant, and vice versa, and hence:

$$W_{if}(k_i) = \Upsilon'' \int_{-\infty}^{+\infty} \int_0^{\infty} \int_0^{\pi/2} \frac{|G_{if}(K_z)|^2}{K_{xy}^2 + K_z^2} \times \left[\frac{\delta(K_{xy} - \alpha_1)}{K_{xy} - \alpha_2} + \frac{\delta(K_{xy} - \alpha_2)}{|K_{xy} - \alpha_1|}\right] K_{xy}\,\mathrm{d}\phi'\,\mathrm{d}K_{xy}\,\mathrm{d}K_z \tag{10.72}$$

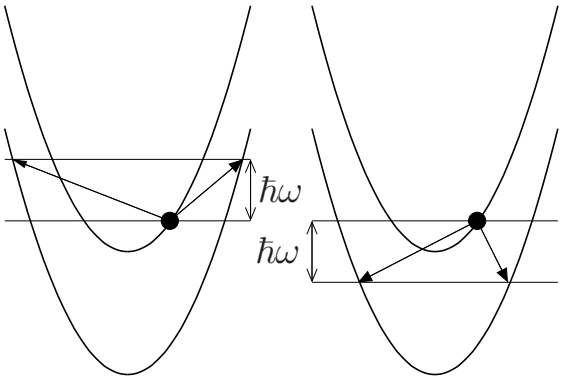


Figure 10.5: The two possible scattering events for (left) phonon absorption and (right) phonon emission

Hence, completing the integration over K_{xy}:

$$W_{if}(k_i) = \Upsilon'' \int_{-\infty}^{+\infty} \int_{0}^{\pi/2} |G_{if}(K_z)|^2 \times \left(\frac{1}{\alpha_1^2 + K_z^2} \frac{\alpha_1}{\alpha_1 - \alpha_2} + \frac{1}{\alpha_2^2 + K_z^2} \frac{\alpha_2}{\alpha_1 - \alpha_2} \right) \mathrm{d}\phi' \, \mathrm{d}K_z \tag{10.73}$$

Recalling the quadratic for K_{xy} in the δ-function argument of equation (10.70):

$$K_{xy} = k_i \cos\phi' \pm \sqrt{k_i^2 \cos^2\phi' - \frac{2m^*\Delta}{\hbar^2}} \tag{10.74}$$

where it is known that as the roots for K_{xy} are real and distinct, then the argument of the square root function is greater than zero. Furthermore, as it has been specified that $\alpha_1 > \alpha_2$:

$$\alpha_1 = k_i \cos\phi' + \sqrt{k_i^2 \cos^2\phi' - \frac{2m^*\Delta}{\hbar^2}} \tag{10.75}$$

and

$$\alpha_2 = k_i \cos\phi' - \sqrt{k_i^2 \cos^2\phi' - \frac{2m^*\Delta}{\hbar^2}} \tag{10.76}$$

Therefore:

$$\alpha_1 + \alpha_2 = 2k_i \cos\phi' \tag{10.77}$$

and

$$\alpha_1\alpha_2 = \frac{2m^*\Delta}{\hbar^2} \tag{10.78}$$

which also come from the 'sum of the roots=$-b/a$' and 'product of the roots=c/a'. In addition:

$$\alpha_1 - \alpha_2 = 2\sqrt{k_i^2 \cos^2\phi' - \frac{2m^*\Delta}{\hbar^2}} \tag{10.79}$$

Hence, if equation (10.73) can be manipulated to only contain the roots in the form of these simple constructions, then a reasonably compact expression may be obtained. Consider the term in parentheses in equation (10.73), i.e.

$$\begin{aligned}\frac{1}{\alpha_1^2+K_z^2}\frac{\alpha_1}{\alpha_1-\alpha_2}+\frac{1}{\alpha_2^2+K_z^2}\frac{\alpha_2}{\alpha_1-\alpha_2} &= \frac{(\alpha_2^2+K_z^2)\alpha_1+(\alpha_1^2+K_z^2)\alpha_2}{(\alpha_1^2+K_z^2)(\alpha_2^2+K_z^2)(\alpha_1-\alpha_2)}\\ &= \frac{\alpha_1\alpha_2^2+\alpha_1K_z^2+\alpha_1^2\alpha_2+\alpha_2K_z^2}{(\alpha_1-\alpha_2)(\alpha_1^2\alpha_2^2+\alpha_1^2K_z^2+\alpha_2^2K_z^2+K_z^4)}\\ &= \frac{(\alpha_1+\alpha_2)(K_z^2+\alpha_1\alpha_2)}{(\alpha_1-\alpha_2)\{(\alpha_1\alpha_2)^2+[(\alpha_1+\alpha_2)^2-2\alpha_1\alpha_2]K_z^2+K_z^4\}}\end{aligned} \tag{10.80}$$

Using the forms for $\alpha_1+\alpha_2$, $\alpha_1\alpha_2$, and $\alpha_1-\alpha_2$ in equations (10.77)–(10.79):

$$\begin{aligned}&\frac{1}{\alpha_1^2+K_z^2}\frac{\alpha_1}{\alpha_1-\alpha_2}+\frac{1}{\alpha_2^2+K_z^2}\frac{\alpha_2}{\alpha_1-\alpha_2}\\ &= \frac{2k_i\cos\phi'\left[K_z^2+\left(\frac{2m^*\Delta}{\hbar^2}\right)\right]}{2\sqrt{k_i^2\cos^2\phi'-\frac{2m^*\Delta}{\hbar^2}}\left[\left(\frac{2m^*\Delta}{\hbar^2}\right)^2+\left(4k_i^2\cos^2\phi'-2\frac{2m^*\Delta}{\hbar^2}\right)K_z^2+K_z^4\right]}\\ &= \frac{k_i\cos\phi'\left[K_z^2+\left(\frac{2m^*\Delta}{\hbar^2}\right)\right]}{\sqrt{k_i^2\cos^2\phi'-\frac{2m^*\Delta}{\hbar^2}}\left[\left(K_z^2-\frac{2m^*\Delta}{\hbar^2}\right)^2+4k_i^2K_z^2\cos^2\phi'\right]}\end{aligned} \tag{10.81}$$

Substituting back into equation (10.73) gives:

$$\begin{aligned}W_{if}(k_i) = \Upsilon''\int_{-\infty}^{+\infty}\int_0^{\pi/2}|G_{if}(K_z)|^2k_i\left[K_z^2+\left(\frac{2m^*\Delta}{\hbar^2}\right)\right]\\ \times\ \frac{\cos\phi'}{\sqrt{k_i^2\cos^2\phi'-\frac{2m^*\Delta}{\hbar^2}}\left[\left(K_z^2-\frac{2m^*\Delta}{\hbar^2}\right)^2+4k_i^2K_z^2\cos^2\phi'\right]}\,\mathrm{d}\phi'\,\mathrm{d}K_z\end{aligned} \tag{10.82}$$

Thus far, the derivation has followed fairly standard methods. However, it is possible to proceed further with the analytical work and evaluate the integral over the angle ϕ' by using the innovative approach of Hagston, Piorek and Harrison [11]. The remainder of this section is dedicated to this procedure, but if only the result is required, then the reader can safely skip to equation (10.111).

For now, though, consider this integral over the angle ϕ', where there is clearly a maximum value for ϕ' which occurs when the argument of the square root function in equation (10.74) becomes zero, which is given by $k_i\cos\phi'_{\max}=\sqrt{2m^*\Delta/\hbar^2}$. This integral takes the form:

$$I=\int_0^{\phi'_{\max}}\frac{\cos\phi'}{\sqrt{a\cos^2\phi'-b(c+d\cos^2\phi')}}\,\mathrm{d}\phi' \tag{10.83}$$

where:

$$a=k_i^2,\qquad b=\frac{2m^*\Delta}{\hbar^2},\qquad c=\left(K_z^2-\frac{2m^*\Delta}{\hbar^2}\right)^2,\qquad d=4k_i^2K_z^2 \tag{10.84}$$

Then:

$$I = \int_0^{\phi'_{\max}} \frac{\sin\phi' \cos\phi'}{\sin\phi' \sqrt{a\cos^2\phi' - b}(c + d\cos^2\phi')} \, \mathrm{d}\phi' \tag{10.85}$$

Putting $x = \cos^2\phi'$, then $\mathrm{d}x = -2\sin\phi'\cos\phi' \, \mathrm{d}\phi'$, and therefore:

$$I = \int_1^{\cos^2\phi'_{\max}} \frac{-\frac{1}{2}}{\sqrt{1-x}\sqrt{ax-b}(c+dx)} \, \mathrm{d}x \tag{10.86}$$

$$\therefore I = \frac{1}{2}\int_{\cos^2\phi'_{\max}}^1 \frac{1}{\sqrt{-ax^2 + (a+b)x - b}(c+dx)} \, \mathrm{d}x \tag{10.87}$$

which is the same as:

$$I = \frac{1}{2\sqrt{a}}\int_{\cos^2\phi'_{\max}}^1 \frac{1}{\sqrt{\frac{(a-b)^2}{4a^2} - \left(x - \frac{a+b}{2a}\right)^2}(c+dx)} \, \mathrm{d}x \tag{10.88}$$

Consider the substitution:

$$x - \frac{a+b}{2a} = \frac{a-b}{2a}\cos\alpha \tag{10.89}$$

and so then:

$$\mathrm{d}x = -\left(\frac{a-b}{2a}\right)\sin\alpha \, \mathrm{d}\alpha \tag{10.90}$$

when $x = \cos^2\phi'_{\max} = b/a$; therefore:

$$\frac{b}{a} - \frac{a+b}{2a} = \frac{a-b}{2a}\cos\alpha \tag{10.91}$$

$$\therefore 2b - a - b = (a-b)\cos\alpha \tag{10.92}$$

i.e. $\cos\alpha = -1$, which implies that the lower limit of the integral becomes π. In addition, when $x = 1$, $\cos\alpha = 1$, which implies that the upper limit is 0. Substituting into equation (10.88) yields:

$$I = \frac{1}{2\sqrt{a}}\int_\pi^0 \frac{-\left(\frac{a-b}{2a}\right)\sin\alpha}{\sqrt{\left(\frac{a-b}{2a}\right)^2 - \left(\frac{a-b}{2a}\cos\alpha\right)^2\left(c + d\left(\frac{a+b}{2a} + \frac{a-b}{2a}\cos\alpha\right)\right)}} \, \mathrm{d}\alpha \tag{10.93}$$

Now the factor $\left(\frac{a-b}{2a}\right)^2$ can be taken out of the square root, and cancelled with the factor in the numerator; the remainder of the argument then becomes $\sqrt{1-\cos^2\alpha}$, which also cancels with the numerator, and hence:

$$I = \frac{1}{2\sqrt{a}}\int_0^\pi \frac{1}{c + d\left(\frac{a+b}{2a} + \frac{a-b}{2a}\cos\alpha\right)} \, \mathrm{d}\alpha \tag{10.94}$$

Writing $e = c + d(a+b)/(2a)$ and $f = d(a-b)/(2a)$, then (for later):

$$e + f = c + d \quad \text{and} \quad e - f = c + d\frac{b}{a} \tag{10.95}$$

Substituting for e and f gives:

$$I = \frac{1}{2\sqrt{a}} \int_0^{\pi} \frac{1}{e + f\cos\alpha}\, \mathrm{d}\alpha \tag{10.96}$$

Consider the substitution $t = \tan(\alpha/2)$. Then:

$$\mathrm{d}t = \frac{1}{2}\sec^2\left(\frac{\alpha}{2}\right)\mathrm{d}\alpha = \frac{1}{2}\left[1 + \tan^2\left(\frac{\alpha}{2}\right)\right]\mathrm{d}\alpha = \frac{1}{2}(1+t^2)\,\mathrm{d}\alpha \tag{10.97}$$

Making use of the tan half-angle formula (see [12], p. 72):

$$\cos\alpha = \frac{1-t^2}{1+t^2} \tag{10.98}$$

and changing the integral limits, equation (10.96) then becomes:

$$I = \frac{1}{2\sqrt{a}} \int_0^{\infty} \frac{1}{e + f\left(\frac{1-t^2}{1+t^2}\right)}\, \frac{2}{1+t^2}\, \mathrm{d}t \tag{10.99}$$

Multiplying the top and bottom of the above equation by $(1+t^2)$ then gives the following:

$$I = \frac{1}{2\sqrt{a}} \int_0^{\infty} \frac{2}{e(1+t^2) + f(1-t^2)}\, \mathrm{d}t = \frac{1}{\sqrt{a}} \int_0^{\infty} \frac{1}{e + f + (e-f)t^2}\, \mathrm{d}t \tag{10.100}$$

This is a standard form, and given that the coefficient of t^2 is greater than zero, the result is then given by the $(\Delta = 4(e+f)(e-f) > 0)$ component of equation (2.172) in [13], i.e.

$$I = \frac{1}{\sqrt{a}} \left[\frac{2}{\sqrt{4(e+f)(e-f)}} \arctan\left(\frac{2(e-f)t}{\sqrt{4(e+f)(e-f)}} \right) \right]_0^{\infty} \tag{10.101}$$

which when evaluated gives:

$$I = \frac{1}{\sqrt{a}} \frac{2}{\sqrt{4(e+f)(e-f)}} \left(\frac{\pi}{2} - 0 \right) \tag{10.102}$$

Recalling the forms for a, b, c and d given in equation (10.84) and substituting these into equations (10.95) then gives the following:

$$e + f = \left(K_z^2 - \frac{2m^*\Delta}{\hbar^2} \right)^2 + 4k_i^2 K_z^2 \tag{10.103}$$

and

$$e - f = \left(K_z^2 - \frac{2m^*\Delta}{\hbar^2} \right)^2 + 4K_z^2 \frac{2m^*\Delta}{\hbar^2} = \left(K_z^2 + \frac{2m^*\Delta}{\hbar^2} \right)^2 \tag{10.104}$$

Substituting for both $e+f$ and $e-f$ in equation (10.102), and recalling that $a = k_i^2$, then:

$$I = \frac{1}{2k_i} \frac{\pi}{\sqrt{\left(K_z^2 - \frac{2m^*\Delta}{\hbar^2}\right)^2 + 4k_i^2 K_z^2}\left(K_z^2 + \frac{2m^*\Delta}{\hbar^2}\right)} \tag{10.105}$$

$$\therefore I = \frac{1}{2k_i} \frac{\pi}{\sqrt{K_z^4 + 2K_z^2\left(2k_i^2 - \frac{2m^*\Delta}{\hbar^2}\right) + \left(\frac{2m^*\Delta}{\hbar^2}\right)^2}\left(K_z^2 + \frac{2m^*\Delta}{\hbar^2}\right)} \tag{10.106}$$

With this analytical form for the integral over the angle ϕ', the original equation (equation (10.82)) becomes:

$$W_{if}(k_i) = \frac{\Upsilon''}{2} \int_{-\infty}^{+\infty} \frac{\pi |G_{if}(K_z)|^2}{\sqrt{K_z^4 + 2K_z^2\left(2k_i^2 - \frac{2m^*\Delta}{\hbar^2}\right) + \left(\frac{2m^*\Delta}{\hbar^2}\right)^2}} \, \mathrm{d}K_z \tag{10.107}$$

This last equation represents the lifetime of a carrier in an initial subband i with *any* in-plane wave vector k_i; strictly speaking only those k_i that satisfy energy conservation can have a lifetime. This information was really lost when the integration over the in-plane phonon wave vector K_{xy} was performed to remove the second (energy conservation) δ-function. It can be put back in with a Heaviside unit step function:

$$\Theta\left(k_i^2 - \frac{2m^*\Delta}{\hbar^2}\right)$$

Recalling that $\Delta = E_f - E_i \mp \hbar\omega$, the Heaviside function ensures that there are only finite lifetimes when:

$$\frac{\hbar^2 k_i}{2m^*} > E_f - E_i \mp \hbar\omega \tag{10.108}$$

Remembering that the upper sign describes phonon absorption, this would imply:

$$E_i + \frac{\hbar^2 k_i}{2m^*} + \hbar\omega > E_f \tag{10.109}$$

and for emission:

$$E_i + \frac{\hbar^2 k_i}{2m^*} > E_f + \hbar\omega \tag{10.110}$$

which are clearly the desired results, hence finally:

$$W_{if}(k_i) = \frac{\Upsilon''}{2}\Theta\left(k_i^2 - \frac{2m^*\Delta}{\hbar^2}\right) \int_{-\infty}^{+\infty} \frac{\pi |G_{if}(K_z)|^2}{\sqrt{K_z^4 + 2K_z^2\left(2k_i^2 - \frac{2m^*\Delta}{\hbar^2}\right) + \left(\frac{2m^*\Delta}{\hbar^2}\right)^2}} \, \mathrm{d}K_z \tag{10.111}$$

The case for a negative Δ follows a similar route and leads to the same end result. Therefore, equation (10.111) represents the final form of the scattering rate of carriers in a subband i with an in-plane wave vector k_i before scattering by an LO phonon. The information regarding whether such an event is with respect to absorption of a phonon or emission, and the final state of the carrier, is incorporated within the variables Δ and P' (which is within Υ'').

This result is a particularly powerful expression because it is applicable to all two-dimensional carrier distributions, regardless of the particular form for the wave functions. Such information is wrapped up in the form factor $G_{if}(K_z)$, and thus the carrier–LO phonon scattering rate can be calculated for any semiconductor heterostructure simply by evaluating a one-dimensional integral.

10.7 Application to conduction subbands

In the previous section, the carrier–LO phonon scattering rate, which is the reciprocal of the lifetime, was derived for a two-dimensional distribution, as found in the subbands formed in quantum well systems. In the present section, this result, as summarised in equation (10.111), will be applied to a variety of examples in order to gain an intuitive understanding of this important phenomenon.

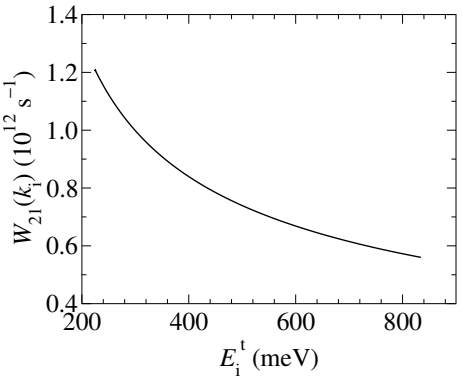


Figure 10.6: The scattering rate via LO phonon emission for an electron initially in the second subband, and finally in the ground state, of a 100 Å GaAs infinitely deep quantum well at a lattice temperature of 77 K

Figure 10.6 displays the intersubband scattering rate as a function of the total initial energy E_i^{t}, as defined in equation (10.59), for an electron in the second subband of an infinitely deep quantum well with respect to LO phonon emission and scattering into the ground state. (Note here the initial carrier energy domain is from the subband minimum upwards.) The scattering rate increases as the carrier approaches the subband minimum, which in this case is around 220 meV. This is a general result and occurs for subband separations greater than the LO phonon energy.

If the quantum well width is allowed to increase, then the energy separation between the initial and final subbands decreases. Figure 10.7 illustrates the effect that such a series of calculations has on the scattering rate. It can be seen that in all cases the scattering rate increases; however, for quantum well widths greater than 300 Å, there is a small region where the scattering rate falls very rapidly to zero. This is indicative of the subband separation, $E_2 - E_1$, being less than the LO phonon energy.

The 'cut-off' in the scattering rate occurs because the electrons in the upper subband do not have sufficient energy to emit an LO phonon and hence are unable to scatter. This feature is illustrated schematically in Fig. 10.8. Moving from the upper right down the curve, electrons have sufficient kinetic energy which, when combined with the potential energy from being in the upper subband, allows them to emit a phonon; however, the third electron represents the minimum kinetic energy for scattering. Below this point, the electrons are less than an LO phonon away from the energy minimum of the *complete* system and hence cannot scatter.

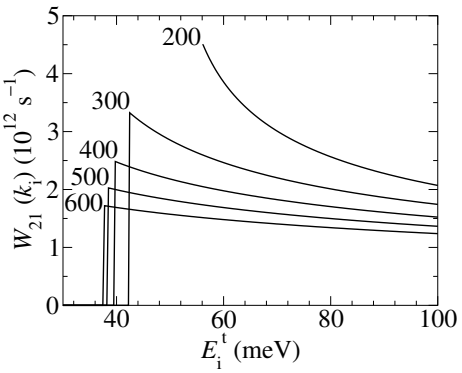


Figure 10.7: The scattering rate as given in Fig. 10.6, but for well widths 200, 300, 400, 500 and 600 Å

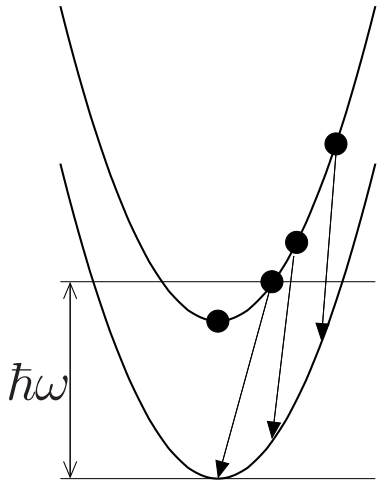


Figure 10.8: The effect of the LO phonon energy 'cut-off' on intersubband electron scattering

Figure 10.9 shows the corresponding intrasubband scattering rate for an electron in the second subband. It can be seen that the behaviour of the rate with the initial energy is qualitatively similar to the intersubband case shown in Fig. 10.6. However, in addition there is a cut-off energy at 260 meV, which is an LO phonon energy above the subband minima. This is always the case for intrasubband scattering—when a carrier is within a phonon energy of the subband minima it cannot emit a phonon. Note that the scattering rate is almost an order of magnitude higher for the intrasubband case than for the intersubband case. This actually represents something of a minimum, as generally the intrasubband rate is between one and two orders of magnitude higher than the intersubband rate. This is because the overlap of the wave function with itself is always complete, whereas the overlap of two distinct wave

functions is often only partial. In this case, with this hypothetical infinitely deep quantum well, the overlap of the wave functions for the intersubband case is higher than the usual situation.

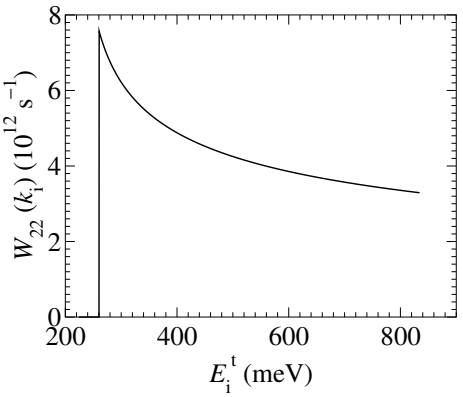


Figure 10.9: The intrasubband scattering rate via LO phonon emission for the same case as shown in Fig. 10.6

10.8 Mean intersubband longitudinal optic phonon scattering rate

The formula shown in equation (10.111) gives the lifetime of a carrier in a particular subband with a definite in-plane wave vector of magnitude k_i with respect to scattering with an LO phonon into another subband. The mean scattering rate from subband i to f can now be found by substituting this into equation (10.28).

Figure 10.10 shows this calculated mean for the same system as before, i.e. a GaAs infinitely deep quantum well, for the case of scattering via the emission of an LO phonon from the second subband to the ground subband. In this series of calculations the width of the well was varied in order to scan the difference between the energy band minima (denoted here by $E_{21} = E_2 - E_1$) through the LO phonon energy, and the electron density in each subband was assumed to be 10^{10} cm^{-2}. The *electron temperature*, i.e. the temperature inserted into the Fermi–Dirac distribution function, was taken equal to the lattice temperature—in systems under excitation where non-equilibrium distributions will be present, this assumption is almost certainly not true. Again, the infinitely deep quantum well is a good illustrative example, as the overlap of the wave functions, which is contained within the form factor G_{if}, does not change, i.e. the effect on the scattering rate is due entirely to the energy separation.

In the long-range scan, shown in Fig. 10.10 (left), it can be seen that, as the subband separation decreases, the scattering rate increases up to almost a 'resonance' point and then decreases rapidly. At energies above the resonance, the scattering rate has only a weak temperature dependence, but below it, the dependence is stronger. Figure 10.10 (right) illustrates this resonance effect more clearly, for a smaller range of energies. As may be

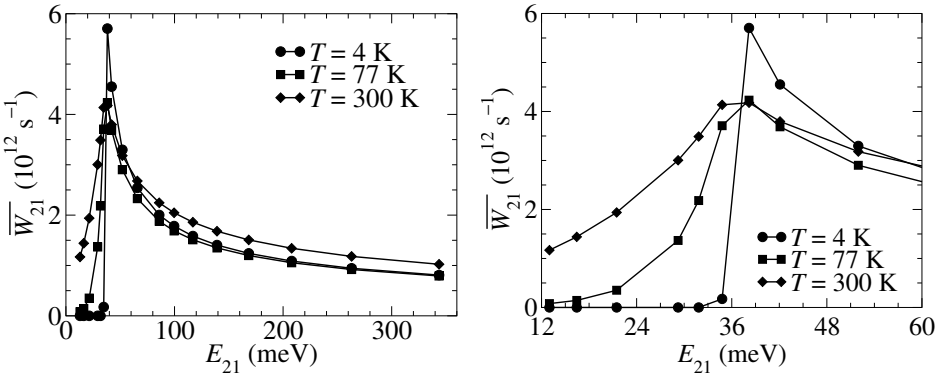


Figure 10.10: The mean scattering rate averaged over distributions in both the initial and final subbands, as a function of the subband separation, shown for three different temperatures

expected for this fixed phonon energy, which in GaAs is $36\,\mathrm{meV}$, the peak in the scattering rate occurs when the subband separation is equal to the phonon energy. The right-hand figure highlights well the strong temperature dependence of the scattering rate for subband separations below the LO phonon energy. At very low temperatures, the 'cut-off' in the scattering is almost as complete as that shown by the single-carrier case in the previous section. However, as the temperature increases, the carrier distributions broaden, so although the subband separation remains below the phonon energy, the carriers in the upper level spread up the subband, with a proportion having enough kinetic energy to be able to emit an LO phonon and scatter to the lower level. As the temperature increases, this proportion increases and hence the mean scattering rate also increases.

10.9 Ratio of emission to absorption

In the previous section, the importance of the carrier energy for emission has been demonstrated. Converse to this, for the case of carrier scattering by absorption of a phonon, it is not the energy of the carrier that is the important issue but rather the number of phonons available—the more phonons, the more likely an absorption process. The phonon density, given in equation (10.29), increases as the temperature rises, thus increasing the probability of an absorption.

Figure 10.11 displays the results of calculations of the ratio of the emission to the electron–LO phonon absorption rate, for the same series of quantum wells as in the previous section. It can be seen that the emission rate is always larger than the absorption rate; thus given a carrier population in an excited subband, when left to reach equilibrium the carriers will always emit more phonons than they absorb and hence scatter down to the ground state.

The linearities of the graphs shown in Fig. 10.11 hint at a relationship of the form:

$$\overline{W_{21}}/\overline{W_{12}} \propto \exp\left(\text{constant} \times E_{21}\right) \tag{10.112}$$

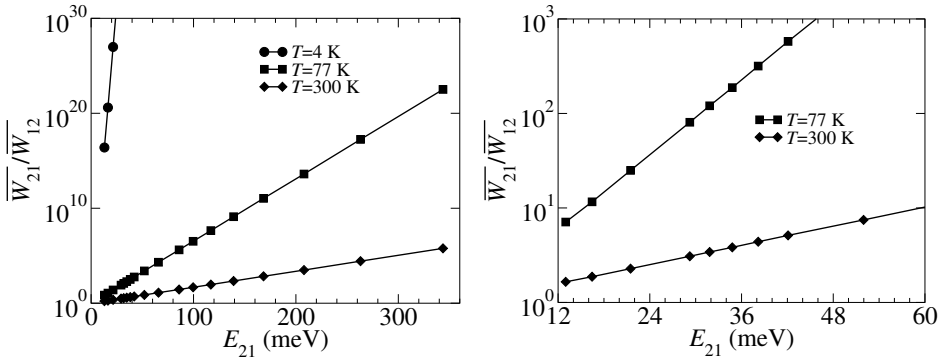


Figure 10.11: The ratio of the mean electron–LO phonon emission ($\overline{W_{21}}$) to absorption ($\overline{W_{12}}$) rate, as a function of the subband separation (left) and a more detailed view around the LO phonon energy (right)

and in fact, a numerical analysis of the data shows that the 'constant' is equal to $1/kT$, i.e.

$$\overline{W_{21}}/\overline{W_{12}} \propto \exp\left(\frac{E_{21}}{kT}\right) \tag{10.113}$$

This simple relationship between the ratio of the emission to the absorption scattering rates, the subband separation and the temperature is helpful in summarising the data presented in Fig. 10.11. For a fixed temperature, this ratio increases as the energy separation between the two levels increases, while for a given subband separation, increasing temperature leads to a decrease in the ratio.

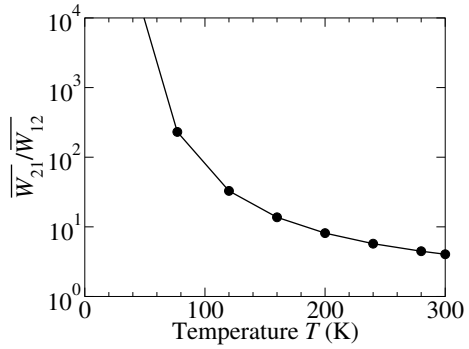


Figure 10.12: The ratio of the mean electron–LO phonon emission ($\overline{W_{21}}$) to absorption ($\overline{W_{12}}$) rate, as a function of temperature T for a fixed subband separation E_{21} equal to the LO phonon energy of $36\,\mathrm{meV}$

These conclusions are not that obvious *a priori*, and go beyond the simple intuitive picture that the ratio of the scattering rates is controlled by the ratio of the phonon densities, i.e. $(N_0 + 1)/N_0$.

For a more detailed investigation into the temperature dependency of the ratio of the emission to absorption scattering rates, consider the data shown in Fig. 10.12, corresponding to the fixed subband separation of E_{21} equal to the LO phonon energy, which in this case is $36\,\mathrm{meV}$. At room temperature, the ratio is equal to the well-known result of 4; however, as the temperature decreases, the ratio of these mean scattering rates increases very rapidly, and by $77\,\mathrm{K}$ emission is more than two orders of magnitude more likely than absorption.

10.10 Screening of the longitudinal optical phonon interaction

The LO phonon interaction is a polar interaction, thus it can be influenced by the presence of other charges. In particular, in a doped semiconductor there can be many charges which are free to move in an electromagnetic field and, like Lenz's law in electromagnetic induction, they move to oppose any change. This idea is known as 'screening' and its effect is to reduce the scattering rate due to LO phonons.

The screening model of Park *et al*. [14] can be implemented in a simple way in the formalism here by making the substitution:

$$K_z^2 \longrightarrow K_z^2 \left(1 + \frac{\lambda_\mathrm{s}^2}{K_z^2}\right)^2 \tag{10.114}$$

in equation (10.111), except in the form factor $G_{if}(K_z)$. The quantity λ_s is known as the 'inverse screening length' and, for systems with a majority carrier type, this would simplify to:

$$\lambda_\mathrm{s}^2 = \frac{e^2}{\pi\hbar^2\epsilon_\mathrm{s}} \sum_j \left\{ \frac{\sqrt{2m^* E_j}\, m^* f^{\mathrm{FD}}(E_j)}{\pi\hbar} \right\} \tag{10.115}$$

where the index j includes all occupied subbands. The unusual parameters employed in Figs. 10.13 and 10.14 were chosen to give direct comparisons with Figs. 1 and 2 of Park *et al*. and the effects of introducing screening are very similar.

The peaks in the curves in Fig. 10.13 occur when the energy separation between the subbands is in resonance with the LO phonon energy, in this case $36\,\mathrm{meV}$. Around this energy, screening at this carrier density can reduce the scattering rate by a factor of 2 or 3, though further from resonance screening has a reduced effect. Figure 10.14 shows the effect of the temperature of the electron distributions in the subbands on the lifetime for two different carrier densities. Such elevated electron temperatures occur commonly in electronic and optoelectronic devices, which are all driven by energy some of which is absorbed directly by the electron gas.

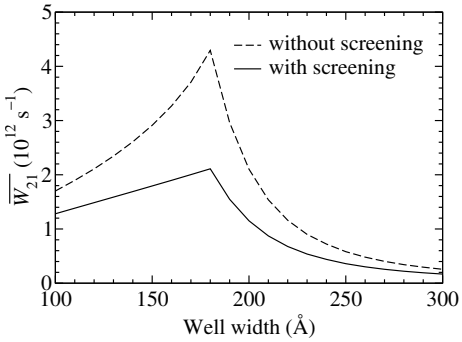


Figure 10.13: The electron–LO phonon scattering rate from the first excited to the ground subband as a function of well width for a GaAs quantum well surrounded by $60\,$Å $Ga_{0.7}Al_{0.3}As$ barriers. The lattice temperature was taken as $15\,$K, the electron temperature as $100\,$K and the carrier density as 10^{11} cm^{-2}

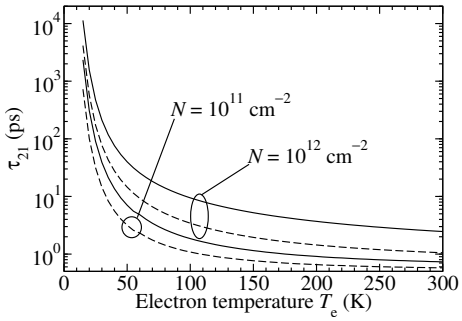


Figure 10.14: The electron–LO phonon scattering rate from the first excited to the ground subband as a function of the temperature of the electron distributions (known as the 'electron temperature') for a $220\,$Å GaAs quantum well surrounded by $60\,$Å $Ga_{0.7}Al_{0.3}As$ barriers. The lattice temperature was taken as $15\,$K

10.11 Acoustic deformation potential scattering

Combining Lundstrom's [8] equations (2.56) and (2.59), summing over all bulk phonon wave functions, which are of the form:

$$\frac{\mathrm{e}^{-i\mathbf{K}_{xy}\bullet\mathbf{r}_{xy}}}{A^{\frac{1}{2}}}\frac{\mathrm{e}^{-iK_{z}\bullet z}}{L^{\frac{1}{2}}} \tag{10.116}$$

and taking a carrier wave function in a heterostructure as before:

$$\psi = \psi(z)\frac{\mathrm{e}^{-i\mathbf{k}_{\bullet}\mathbf{r}_{xy}}}{A^{\frac{1}{2}}} \tag{10.117}$$

gives on insertion into Fermi's golden rule:

$$W_{if}(k_i) = \frac{2\pi}{\hbar}\sum_{\mathbf{k}_f}\frac{D_{\mathrm{A}}^2\hbar}{2\rho\omega_{\mathrm{s}}}\left(N_0+\frac{1}{2}\mp\frac{1}{2}\right)$$

$$\times\left|\iint\psi_f^*(z)\frac{\mathrm{e}^{i\mathbf{k}_f\bullet\mathbf{r}_{xy}}}{A^{\frac{1}{2}}}\sum_{\mathbf{K}_{xy}}\sum_{K_z}K\frac{\mathrm{e}^{-i\mathbf{K}_{xy}\bullet\mathbf{r}_{xy}}}{A^{\frac{1}{2}}}\frac{\mathrm{e}^{-iK_zz}}{L^{\frac{1}{2}}}\psi_i(z)\frac{\mathrm{e}^{-i\mathbf{k}_i\bullet\mathbf{r}_{xy}}}{A^{\frac{1}{2}}}\mathrm{d}z\ \mathrm{d}\mathbf{r}_{xy}\right|^2$$

$$\times\delta(E_f^{\mathrm{t}}-E_i^{\mathrm{t}}\mp\hbar\omega_{\mathrm{s}}) \tag{10.118}$$

where $\hbar\omega_{\mathrm{s}}$ is the energy of the acoustic phonon. This then gives:

$$W_{if}(k_i) = \frac{2\pi}{\hbar}\sum_{\mathbf{k}_f}\frac{D_{\mathrm{A}}^2\hbar}{2\rho\omega_{\mathrm{s}}}\left(N_0+\frac{1}{2}\mp\frac{1}{2}\right)$$

$$\times\left|\sum_{\mathbf{K}_{xy}}\sum_{K_z}\int\psi_f^*(z)\frac{\mathrm{e}^{-iK_zz}}{L^{\frac{1}{2}}}\psi_i(z)\ \mathrm{d}zK\int\frac{\mathrm{e}^{i\mathbf{k}_f\bullet\mathbf{r}_{xy}}}{A^{\frac{1}{2}}}\frac{\mathrm{e}^{-i\mathbf{K}_{xy}\bullet\mathbf{r}_{xy}}}{A^{\frac{1}{2}}}\frac{\mathrm{e}^{-i\mathbf{k}_i\bullet\mathbf{r}_{xy}}}{A^{\frac{1}{2}}}\ \mathrm{d}\mathbf{r}_{xy}\right|^2$$

$$\times\delta(E_f^{\mathrm{t}}-E_i^{\mathrm{t}}\mp\hbar\omega_{\mathrm{s}}) \tag{10.119}$$

Now the integral over z is just the form factor (see equation (10.47)), hence:

$$W_{if}(k_i) = \frac{2\pi}{\hbar}\sum_{\mathbf{k}_f}\frac{D_{\mathrm{A}}^2\hbar}{2\rho\omega_{\mathrm{s}}}\left(N_0+\frac{1}{2}\mp\frac{1}{2}\right)$$

$$\times\left|\sum_{\mathbf{K}_{xy}}\sum_{K_z}\frac{G_{if}(K_z)}{L^{\frac{1}{2}}}K\int\frac{\mathrm{e}^{-i(\mathbf{k}_i-\mathbf{k}_f+\mathbf{K}_{xy})\bullet\mathbf{r}_{xy}}}{A^{\frac{3}{2}}}\mathrm{d}\mathbf{r}_{xy}\right|^2\delta(E_f^{\mathrm{t}}-E_i^{\mathrm{t}}\mp\hbar\omega_{\mathrm{s}}) \tag{10.120}$$

Converting the integral over the x–y plane (denoted by $\mathrm{d}\mathbf{r}_{xy}$) into a δ-function gives a factor of 2π per dimension, i.e.

$$W_{if}(k_i) = \frac{2\pi}{\hbar}\sum_{\mathbf{k}_f}\frac{D_{\mathrm{A}}^2\hbar}{2\rho\omega_{\mathrm{s}}LA^3}\left(N_0+\frac{1}{2}\mp\frac{1}{2}\right)$$

$$\times\left|\sum_{\mathbf{K}_{xy}}\sum_{K_z}G_{if}(K_z)K(2\pi)^2\delta\left(\mathbf{k}_i-\mathbf{k}_f+\mathbf{K}_{xy}\right)\right|^2\delta(E_f^{\mathrm{t}}-E_i^{\mathrm{t}}\mp\hbar\omega_{\mathrm{s}}) \tag{10.121}$$

Now converting the summations over K_z and $\mathbf{K}_{xy}$ into integrals introduces factors of $L/(2\pi)$ and $A/(2\pi)^2$ respectively, giving:

$$W_{if}(k_i) = \frac{2\pi}{\hbar}\sum_{\mathbf{k}_f}\frac{D_{\mathrm{A}}^2\hbar}{2\rho\omega_{\mathrm{s}}LA^3}\left(N_0+\frac{1}{2}\mp\frac{1}{2}\right)$$

$$\times \left| \iint \frac{L}{2\pi} G_{if}(K_z) K \frac{A}{(2\pi)^2} (2\pi)^2 \delta\left(\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}_{xy}\right)\, \mathrm{d}\mathbf{K}_{xy}\, \mathrm{d}K_z \right|^2 \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega_{\mathrm{s}}) \tag{10.122}$$

Just simplifying the constants, then:

$$W_{if}(k_i) = \sum_{\mathbf{k}_f} \frac{D_{\mathrm{A}}^2}{2\rho\omega_{\mathrm{s}} A} \frac{L}{2\pi} \left(N_0 + \frac{1}{2} \mp \frac{1}{2} \right)$$

$$\times \left| \iint G_{if}(K_z) K \delta\left(\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}_{xy}\right)\, \mathrm{d}\mathbf{K}_{xy}\, \mathrm{d}K_z \right|^2 \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega_{\mathrm{s}}) \tag{10.123}$$

Approximating the acoustic branch of the phonon dispersion curve by a linear function, the angular frequency of a sound wave is given by:

$$\omega_{\mathrm{s}} = v_{\mathrm{s}} K \tag{10.124}$$

Hence as ω_{s} is dependent upon the phonon wave vector, it needs to be included within the integrals:

$$W_{if}(k_i) = \sum_{\mathbf{k}_f} \frac{D_{\mathrm{A}}^2}{2\rho v_{\mathrm{s}} A} \frac{L}{2\pi} \left(N_0 + \frac{1}{2} \mp \frac{1}{2} \right)$$

$$\times \left| \iint G_{if}(K_z) \sqrt{K} \delta\left(\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}_{xy}\right)\, \mathrm{d}\mathbf{K}_{xy}\, \mathrm{d}K_z \right|^2 \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega_{\mathrm{s}}) \tag{10.125}$$

Converting the sum over final carrier momentum states $\mathbf{k}_f$ into an integral introduces a factor of $A/(2\pi)^2$, thus:

$$W_{if}(k_i) = \frac{D_{\mathrm{A}}^2}{2\rho v_{\mathrm{s}} A} \frac{L}{2\pi} \frac{A}{(2\pi)^2} \left(N_0 + \frac{1}{2} \mp \frac{1}{2} \right)$$

$$\times \int \left| \iint G_{if}(K_z) \sqrt{K} \delta\left(\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}_{xy}\right)\, \mathrm{d}\mathbf{K}_{xy}\, \mathrm{d}K_z \right|^2 \mathrm{d}\mathbf{k}_f\, \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega_{\mathrm{s}}) \tag{10.126}$$

Expanding the square:

$$W_{if}(k_i) = \frac{D_{\mathrm{A}}^2}{2\rho v_{\mathrm{s}} (2\pi)^2} \frac{L}{2\pi} \left(N_0 + \frac{1}{2} \mp \frac{1}{2} \right)$$

$$\times \int \left(\iint G_{if}(K_z) \sqrt{(K_{xy}^2 + K_z^2)^{\frac{1}{2}}} \delta\left(\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}_{xy}\right)\, \mathrm{d}\mathbf{K}_{xy}\, \mathrm{d}K_z \right)$$

$$\times \left(\iint G_{if}(K_z') \sqrt{({K_{xy}'}^2 + {K_z'}^2)^{\frac{1}{2}}} \delta\left(\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}_{xy}'\right)\, \mathrm{d}\mathbf{K}_{xy}'\, \mathrm{d}K_z' \right)$$

$$\times \mathrm{d}\mathbf{k}_f\, \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega_{\mathrm{s}}) \tag{10.127}$$

Performing the integration over $\mathbf{K}'_{xy}$ first, then the second δ-function implies that there is only a contribution when $\mathbf{K}'_{xy} = -\mathbf{k}_i + \mathbf{k}_f$:

$$W_{if}(k_i) = \frac{D_{\mathrm{A}}^2}{2\rho v_{\mathrm{s}}(2\pi)^2}\frac{L}{2\pi}\left(N_0 + \frac{1}{2} \mp \frac{1}{2}\right) \times \int \left(\iint G_{if}(K_z)\sqrt{(K_{xy}^2 + K_z^2)^{\frac{1}{2}}}\,\delta\left(\mathbf{k}_i - \mathbf{k}_f + \mathbf{K}_{xy}\right)\,\mathrm{d}\mathbf{K}_{xy}\,\mathrm{d}K_z\right) \times \left(\int G_{if}(K'_z)\sqrt{(|-\mathbf{k}_i + \mathbf{k}_f|^2 + K_z'^2)^{\frac{1}{2}}}\,\mathrm{d}K'_z\right)\,\mathrm{d}\mathbf{k}_f\;\delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega_{\mathrm{s}}) \tag{10.128}$$

Now performing the integral over the final carrier momentum $\mathbf{k}_f$, the first δ-function limits the integrals to contributions when $\mathbf{k}_f = \mathbf{k}_i + \mathbf{K}_{xy}$, i.e.

$$W_{if}(k_i) = \frac{D_{\mathrm{A}}^2}{2\rho v_{\mathrm{s}}(2\pi)^2}\frac{L}{2\pi}\left(N_0 + \frac{1}{2} \mp \frac{1}{2}\right) \times \left(\iint G_{if}(K_z)\sqrt{(K_{xy}^2 + K_z^2)^{\frac{1}{2}}}\,\mathrm{d}\mathbf{K}_{xy}\,\mathrm{d}K_z\right) \times \left(\int G_{if}(K'_z)\sqrt{(|-\mathbf{k}_i + \mathbf{k}_i + \mathbf{K}_{xy}|^2 + K_z'^2)^{\frac{1}{2}}}\,\mathrm{d}K'_z\right)\;\delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega_{\mathrm{s}}) \tag{10.129}$$

and therefore:

$$W_{if}(k_i) = \frac{D_{\mathrm{A}}^2}{2\rho v_{\mathrm{s}}(2\pi)^2}\frac{L}{2\pi}\left(N_0 + \frac{1}{2} \mp \frac{1}{2}\right) \times \iiint G_{if}(K_z)G_{if}(K'_z)\sqrt{(K_{xy}^2 + K_z^2)^{\frac{1}{2}}}\sqrt{(K_{xy}^2 + K_z'^2)^{\frac{1}{2}}}\,\mathrm{d}K_z\,\mathrm{d}K'_z\,\mathrm{d}\mathbf{K}_{xy} \times\delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega_{\mathrm{s}}) \tag{10.130}$$

Utilising the same mathematical trick as before (see Section 10.6) allows the mathematics to be forced through to yield expressions that agree with those quoted, but again it is worth reiterating the statement that maybe the scattering rate should be a 'square of an integral' rather than an 'integral of a square'. The procedure equates K'_z to K_z and introduces a factor of $2\pi/L$, thus giving:

$$W_{if}(k_i) = \frac{D_{\mathrm{A}}^2}{2\rho v_{\mathrm{s}}(2\pi)^2}\left(N_0 + \frac{1}{2} \mp \frac{1}{2}\right) \times \iint \left(G_{if}(K_z)\right)^2 (K_{xy}^2 + K_z^2)^{\frac{1}{2}}\,\mathrm{d}K_z\;\,\mathrm{d}\mathbf{K}_{xy}\,\delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega_{\mathrm{s}}) \tag{10.131}$$

Now, as before, the initial and final total energies of the carrier are given by the sum of the subband minima and the kinetic energy. Hence substituting into equation (10.131):

$$W_{if}(k_i) = \frac{D_{\mathrm{A}}^2}{2\rho v_{\mathrm{s}}(2\pi)^2}\left(N_0 + \frac{1}{2} \mp \frac{1}{2}\right)$$

$$\times \iint \left(G_{if}(K_z)\right)^2 \left(K_{xy}^2 + K_z^2\right)^{\frac{1}{2}} \delta\left(E_f + \frac{\hbar^2 k_f^2}{2m^*} - E_i - \frac{\hbar^2 k_i^2}{2m^*} \mp \hbar\omega_\mathrm{s}\right) \mathrm{d}K_z \, \mathrm{d}\mathbf{K}_{xy} \tag{10.132}$$

Following the prescription in Fig. 10.4 and defining the angle between the initial and final carrier momentum states as ϕ, and applying the cosine rule gives:

$$k_f^2 = k_i^2 + K_{xy}^2 - 2k_i K_{xy} \cos\left(\pi - \phi\right) \tag{10.133}$$

$$\therefore k_f^2 = k_i^2 + K_{xy}^2 + 2k_i K_{xy} \cos\phi \tag{10.134}$$

Hence, substituting into equation (10.132):

$$W_{if}(k_i) = \frac{D_\mathrm{A}^2}{2\rho v_\mathrm{s}(2\pi)^2} \left(N_0 + \frac{1}{2} \mp \frac{1}{2}\right) \iint \left(G_{if}(K_z)\right)^2 \left(K_{xy}^2 + K_z^2\right)^{\frac{1}{2}}$$

$$\times\delta\left(E_f + \frac{\hbar^2}{2m^*}\left(k_i^2 + K_{xy}^2 + 2k_i K_{xy}\cos\phi\right) - E_i - \frac{\hbar^2 k_i^2}{2m^*} \mp \hbar\omega_\mathrm{s}\right) \mathrm{d}K_z \, \mathrm{d}\mathbf{K}_{xy} \tag{10.135}$$

Now the energies of phonons from the acoustic branch are generally very small, i.e. a few millielectronvolts, which is small when compared to typical intersubband separations $\Delta E = E_f - E_i$, which range from several tens to several hundred millielectronvolts, hence in this formalism the phonon energy $\hbar\omega_\mathrm{s}$ will be approximated as zero. The physical implication of this is that, *from the viewpoint of the calculation of scattering rates*, the acoustic deformation potential scattering is assumed elastic.

Implementing this and taking the factor $\hbar^2/2m$ out of the δ-function:

$$W_{if}(k_i) = \frac{D_\mathrm{A}^2 m^*}{\rho v_\mathrm{s}(2\pi)^2\hbar^2} \left(N_0 + \frac{1}{2} \mp \frac{1}{2}\right) \iint \left(G_{if}(K_z)\right)^2 \left(K_{xy}^2 + K_z^2\right)^{\frac{1}{2}}$$

$$\times\delta\left(K_{xy}^2 + 2k_i K_{xy}\cos\phi + \frac{2m^*\Delta E}{\hbar^2}\right) \mathrm{d}K_z \, \mathrm{d}\mathbf{K}_{xy} \tag{10.136}$$

Transforming the integral over the in-plane phonon wave vector $\mathbf{K}_{xy}$ into polar coordinates gives:

$$W_{if}(k_i) = \frac{D_\mathrm{A}^2 m^*}{\rho v_\mathrm{s}(2\pi)^2\hbar^2} \left(N_0 + \frac{1}{2} \mp \frac{1}{2}\right) \int \int_0^{2\pi} \int \left(G_{if}(K_z)\right)^2 \left(K_{xy}^2 + K_z^2\right)^{\frac{1}{2}}$$

$$\times\delta\left(K_{xy}^2 + 2k_i K_{xy}\cos\phi + \frac{2m^*\Delta E}{\hbar^2}\right) K_{xy} \, \mathrm{d}K_{xy} \, \mathrm{d}\phi \, \mathrm{d}K_z \tag{10.137}$$

The argument within the δ-function can be factorised—as K_{xy} is a magnitude then solutions for it must be positive, hence the argument can be factorised as $(K_{xy} - \alpha_1)(K_{xy} - \alpha_2)$, where the roots α_1 and α_2 are given by:

$$\alpha_{1,2} = -k_i \cos\phi \pm \sqrt{k_i^2 \cos^2\phi - \frac{2m^*\Delta E}{\hbar^2}} \tag{10.138}$$

where clearly $\alpha_1 > \alpha_2$. Equation (10.137) can therefore be written as:

$$W_{if}(k_i) = \frac{D_{\mathrm{A}}^2 m^*}{\rho v_{\mathrm{s}} (2\pi)^2 \hbar^2} \left(N_0 + \frac{1}{2} \mp \frac{1}{2} \right) \int_0^\infty \int_0^{2\pi} \int_0^\infty \left(G_{if}(K_z) \right)^2 \left(K_{xy}^2 + K_z^2 \right)^{\frac{1}{2}} \times \delta\left((K_{xy} - \alpha_1)(K_{xy} - \alpha_2) \right) \; K_{xy} \, \mathrm{d}K_{xy} \, \mathrm{d}\phi \, \mathrm{d}K_z \tag{10.139}$$

Consider now the integration over K_{xy}. There are only two contributions to the integral, one when K_{xy} is around α_1 and the other when K_{xy} is around α_2. Thus the δ-function can be split into two components:

$$W_{if}(k_i) = \frac{D_{\mathrm{A}}^2 m^*}{\rho v_{\mathrm{s}} (2\pi)^2 \hbar^2} \left(N_0 + \frac{1}{2} \mp \frac{1}{2} \right) \int_0^\infty \int_0^{2\pi} \int_0^\infty \left(G_{if}(K_z) \right)^2 \left(K_{xy}^2 + K_z^2 \right)^{\frac{1}{2}} \times \left[\frac{1}{|K_{xy} - \alpha_2|} \delta\left(K_{xy} - \alpha_1 \right) + \frac{1}{|K_{xy} - \alpha_1|} \delta\left(K_{xy} - \alpha_2 \right) \right] \; K_{xy} \, \mathrm{d}K_{xy} \, \mathrm{d}\phi \, \mathrm{d}K_z \tag{10.140}$$

and now actually performing the integration over K_{xy} gives:

$$W_{if}(k_i) = \frac{D_{\mathrm{A}}^2 m^*}{\rho v_{\mathrm{s}} (2\pi)^2 \hbar^2} \left(N_0 + \frac{1}{2} \mp \frac{1}{2} \right) \int_0^\infty \int_0^{2\pi} \left(G_{if}(K_z) \right)^2 \times \left(\frac{\Theta\left(\alpha_1 \right) \alpha_1 \sqrt{\alpha_1^2 + K_z^2}}{|\alpha_1 - \alpha_2|} + \frac{\Theta\left(\alpha_2 \right) \alpha_2 \sqrt{\alpha_2^2 + K_z^2}}{|\alpha_2 - \alpha_1|} \right) \mathrm{d}\phi \, \mathrm{d}K_z \tag{10.141}$$

where the Heaviside functions ensure there are only contributions for positive α_1 and α_2. Recalling $\alpha_1 > \alpha_2$ then, finally:

$$W_{if}(k_i) = \frac{D_{\mathrm{A}}^2 m^*}{\rho v_{\mathrm{s}} (2\pi)^2 \hbar^2} \left(N_0 + \frac{1}{2} \mp \frac{1}{2} \right) \int_0^\infty \int_0^{2\pi} \left(G_{if}(K_z) \right)^2 \times \left(\frac{\Theta(\alpha_1) \alpha_1 \sqrt{\alpha_1^2 + K_z^2} + \Theta(\alpha_2) \alpha_2 \sqrt{\alpha_2^2 + K_z^2}}{\alpha_1 - \alpha_2} \right) \mathrm{d}\phi \, \mathrm{d}K_z \tag{10.142}$$

which is the same result as that of Piorek [11].

10.12 Application to conduction subbands

Figure 10.15(a) shows the effect of the energy separation on the intersubband scattering rate due to acoustic phonon emission. It can be seen that, for these thermalised distributions, the scattering rate is almost linear. A comparison with the equivalent electron–electron scattering rate, as shown later in Figs. 10.23 and 10.24(a), shows that acoustic deformation potential scattering is much slower (an order of magnitude or more) than electron–electron scattering for small (less than $20\,\mathrm{meV}$) intersubband energy separations at the same carrier density. However, it can be comparable or faster at larger (greater than $100\,\mathrm{meV}$) energy separations. Figure 10.15(b) compares the emission rate given in (a) with the absorption rate (shown as

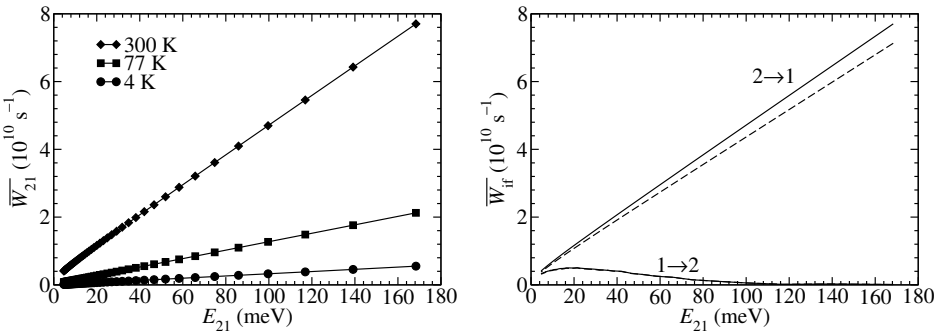


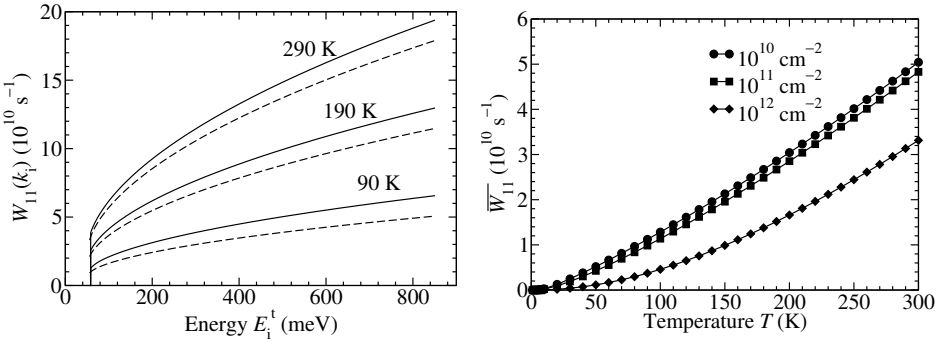
Figure 10.15: The intersubband acoustic deformation potential scattering rates as a function of the subband energy separation $E_{21} = E_2 - E_1$ for an infinitely deep quantum well with electron densities of $1 \times 10^{10}\,\mathrm{cm}^{-2}$ in each of the lowest two subbands; (a) shows the effect of temperature on the scattering rate from the second ($n = 2$) to the first ($n = 1$) subband due to the emission of acoustic phonons, and (b) compares the scattering rates from the second to the first and the first to the second subbands due to emission (solid lines) and absorption (dashed lines) of acoustic phonons at the fixed temperature of $300\,\mathrm{K}$

dashed lines). It can be seen that, for the same pair of initial (2) and final (1) states, the absorption rate remains a little below the emission rate due to the ratio of $N_0 + 1$ to N_0. In contrast to this the emission and absorption rates for carriers to scatter from the ground state (1) to the first excited state (2), which are visible in the lower section of Fig. 10.15(b), reach a maximum and then decrease as the separation between the initial and final states increases. This is to be expected as the acoustic phonon can only supply a few millielectronvolts of energy and as the carriers in the initial state are thermalised most of them are concentrated near the subband minimum and simply cannot gain enough energy to transfer into the higher state.

Figure 10.16(a) shows the microscopic effect of acoustic deformation potential scattering on carriers *within* a subband. In particular, it can be seen that the rate for emission of acoustic phonons is greater than that for absorption for all initial carrier energies and temperatures, and this is why carriers which have a lot of kinetic energy within a subband lose this energy to the lattice and eventually form a thermalised distribution.[2] Note that the presence of the subband minimum produces the 'cut-off' in the data at energies just larger than $50\,\mathrm{meV}$.

Figure 10.16(b) illustrates the results of calculations of the effect of temperature and carrier density on the *thermally averaged* acoustic deformation potential scattering rates. Increasing temperature implies the presence of more phonons and hence the scattering rate increases, while the variations due to the carrier density changes are much smaller and are due to final-state blocking. It is interesting to note that, for (b), the thermally averaged phonon absorption rates (which are not shown) are almost identical to the emission rates under this

[2]This result helps justify the assumption which gives the thermal averaging of carriers, used so often in this chapter, some physical meaning.



(a) (b)

Figure 10.16: (a) The intrasubband acoustic deformation potential scattering rates due to phonon emission (solid lines) and phonon absorption (dashed lines) as a function of the total energy E_i^{t} of the carrier in the initial state (in this case the subband index $i = 1$) in a 100 Å GaAs infinitely deep quantum well with a carrier density of $1 \times 10^{10}\,\mathrm{cm}^{-2}$, and (b) the thermally averaged intrasubband scattering rates due to acoustic phonon emission again in the ground state of a 100 Å GaAs infinitely deep quantum well as a function of the temperature and for several different carrier densities

assumption of thermalised (Fermi–Dirac) carrier distributions; this is a reflection of the idea of an equilibrium.

10.13 Optical deformation potential scattering

Taking the optical deformation potential as a sum over all phonon wave vectors with the interaction term of in Lundstrom's [8] equation (2.67), then in analogy with equation (10.118) for acoustic deformation potential scattering, Fermi's golden rule gives:

$$W_{if}(k_i) = \frac{2\pi}{\hbar} \sum_{\mathbf{k}_f} \frac{D_{\mathrm{o}}^2 \hbar}{2\rho\omega_{\mathrm{o}}} \left(N_0 + \frac{1}{2} \mp \frac{1}{2} \right)$$

$$\times \left| \iint \psi_f^*(z) \frac{\mathrm{e}^{i\mathbf{k}_f \bullet \mathbf{r}_{xy}}}{A^{\frac{1}{2}}} \sum_{\mathbf{K}_{xy}} \sum_{K_z} \frac{\mathrm{e}^{-i\mathbf{K}_{xy} \bullet \mathbf{r}_{xy}}}{A^{\frac{1}{2}}} \frac{\mathrm{e}^{-iK_z z}}{L^{\frac{1}{2}}} \psi_i(z) \frac{\mathrm{e}^{-i\mathbf{k}_i \bullet \mathbf{r}_{xy}}}{A^{\frac{1}{2}}} \mathrm{d}z \; \mathrm{d}\mathbf{r}_{xy} \right|^2$$

$$\times \delta(E_f^{\mathrm{t}} - E_i^{\mathrm{t}} \mp \hbar\omega_{\mathrm{o}}) \tag{10.143}$$

The derivation can now be performed in a very similar manner to that of acoustic deformation potential scattering, as in Section 10.11. However, the optical phonon frequency ω_{o} may be taken as a constant rather than substituting the linear acoustic phonon dispersion as in equation (10.124). As before, the total carrier energies may be found by assuming parabolic

subbands. The resulting expression, in analogy to equation (10.135), is:

$$W_{if}(k_i) = \frac{D_{\mathrm{o}}^2}{2\rho\omega_{\mathrm{o}}(2\pi)^2}\left(N_0 + \frac{1}{2} \mp \frac{1}{2}\right)\iint \left(G_{if}(K_z)\right)^2$$

$$\times\delta\left(E_f + \frac{\hbar^2}{2m^*}\left(\mathbf{k}_i^2 + \mathbf{K}_{xy}^2 + 2k_iK_{xy}\cos\phi\right) - E_i - \frac{\hbar^2\mathbf{k}_i^2}{2m^*} \mp \hbar\omega_{\mathrm{o}}\right)\,\mathrm{d}K_z\,\mathrm{d}\mathbf{K}_{xy} \quad (10.144)$$

Now the energies of phonons from the optical branch are not small, hence the phonon energy cannot be approximated to zero and the collision considered elastic, rather the term $\mp\hbar\omega_{\mathrm{o}}$ has to be retained. Labelling $E_f - E_i \mp \hbar\omega_{\mathrm{o}}$ as Δ, then

$$W_{if}(k_i) = \frac{D_{\mathrm{o}}^2}{2\rho\omega_{\mathrm{o}}(2\pi)^2}\left(N_0 + \frac{1}{2} \mp \frac{1}{2}\right)\iint \left(G_{if}(K_z)\right)^2$$

$$\times\delta\left(\frac{\hbar^2K_{xy}^2}{2m^*} + \frac{\hbar^2k_iK_{xy}\cos\phi}{m^*} + \Delta\right)\,\mathrm{d}K_z\,\mathrm{d}\mathbf{K}_{xy} \quad (10.145)$$

Taking the factor $\hbar^2/(2m^*)$ out of the δ-function:

$$W_{if}(k_i) = \frac{D_{\mathrm{o}}^2m^*}{\rho\omega_{\mathrm{o}}(2\pi)^2\hbar^2}\left(N_0 + \frac{1}{2} \mp \frac{1}{2}\right)\iint \left(G_{if}(K_z)\right)^2 \quad (10.146)$$

$$\times\delta\left(K_{xy}^2 + 2k_iK_{xy}\cos\phi + \frac{2m^*\Delta}{\hbar^2}\right)\,\mathrm{d}K_z\,\mathrm{d}\mathbf{K}_{xy}$$

Transforming the integral over the in-plane phonon wave vector $\mathbf{K}_{xy}$ into plane polar coordinates:

$$W_{if}(k_i) = \frac{D_{\mathrm{o}}^2m^*}{\rho\omega_{\mathrm{o}}(2\pi)^2\hbar^2}\left(N_0 + \frac{1}{2} \mp \frac{1}{2}\right)\int\int_0^{2\pi}\int \left(G_{if}(K_z)\right)^2$$

$$\times\delta\left(K_{xy}^2 + 2k_iK_{xy}\cos\phi + \frac{2m^*\Delta}{\hbar^2}\right)\,K_{xy}\,\mathrm{d}K_{xy}\,\mathrm{d}\phi\,\mathrm{d}K_z \quad (10.147)$$

Again, following previous procedures, consider the integral over K_{xy}. There is only a contribution to this integral when the argument within the δ-function is zero. Now clearly solutions for K_{xy} must be positive (since K_{xy} is the *length* of the in-plane phonon wave vector), hence the argument of the δ-function can be factorised as $(K_{xy} - \alpha_1)(K_{xy} - \alpha_2)$, where again the roots are given by:

$$\alpha_{1,2} = -k_i\cos\phi \pm \sqrt{k_i^2\cos^2\phi - \frac{2m^*\Delta}{\hbar^2}} \quad (10.148)$$

and again $\alpha_1 > \alpha_2$. Equation (10.147) can therefore be written:

$$W_{if}(k_i) = \frac{D_{\mathrm{o}}^2m^*}{\rho\omega_{\mathrm{o}}(2\pi)^2\hbar^2}\left(N_0 + \frac{1}{2} \mp \frac{1}{2}\right)\int\int_0^{2\pi}\int \left(G_{if}(K_z)\right)^2$$

$$\times\delta\left((K_{xy} - \alpha_1)(K_{xy} - \alpha_2)\right)\,K_{xy}\,\mathrm{d}K_{xy}\,\mathrm{d}\phi\,\mathrm{d}K_z \quad (10.149)$$

Again following the same procedure as for acoustic deformation potential scattering, then:

$$W_{if}(k_i) = \frac{D_o^2 m^*}{\rho\omega_o(2\pi)^2\hbar^2}\left(N_0 + \frac{1}{2} \mp \frac{1}{2}\right)\int_0^\infty \int_0^{2\pi} \int_0^\infty (G_{if}(K_z))^2$$

$$\times \left[\frac{1}{|K_{xy} - \alpha_2|}\delta(K_{xy} - \alpha_1) + \frac{1}{|K_{xy} - \alpha_1|}\delta(K_{xy} - \alpha_2)\right] K_{xy}\, \mathrm{d}K_{xy}\, \mathrm{d}\phi\, \mathrm{d}K_z \quad (10.150)$$

and now performing the integration over K_{xy}:

$$W_{if}(k_i) = \frac{D_o^2 m^*}{\rho\omega_o(2\pi)^2\hbar^2}\left(N_0 + \frac{1}{2} \mp \frac{1}{2}\right)\int_0^\infty \int_0^{2\pi} (G_{if}(K_z))^2$$

$$\times \left[\frac{\alpha_1}{|\alpha_1 - \alpha_2|} + \frac{\alpha_2}{|\alpha_2 - \alpha_1|}\right] \mathrm{d}\phi\, \mathrm{d}K_z \quad (10.151)$$

and with $\alpha_1 > \alpha_2$:

$$W_{if}(k_i) = \frac{D_o^2 m^*}{\rho\omega_o(2\pi)^2\hbar^2}\left(N_0 + \frac{1}{2} \mp \frac{1}{2}\right)\int_0^\infty \int_0^{2\pi} (G_{if}(K_z))^2 \frac{\alpha_1 + \alpha_2}{\alpha_1 - \alpha_2}\, \mathrm{d}\phi\, \mathrm{d}K_z \quad (10.152)$$

But:

$$\alpha_1 + \alpha_2 = -2k_i\cos\phi \quad \text{and} \quad \alpha_1 - \alpha_2 = 2\sqrt{k_i^2\cos^2\phi - \frac{2m^*\Delta}{\hbar^2}} \quad (10.153)$$

Therefore:

$$W_{if}(k_i) = -\frac{D_o^2 m^*}{\rho\omega_o(2\pi)^2\hbar^2}\left(N_0 + \frac{1}{2} \mp \frac{1}{2}\right) \quad (10.154)$$

$$\times \int_0^\infty \int_0^{2\pi} (G_{if}(K_z))^2 \frac{k_i\cos\phi}{\sqrt{k_i^2\cos^2\phi - \frac{2m^*\Delta}{\hbar^2}}}\, \mathrm{d}\phi\, \mathrm{d}K_z$$

and as these integrals are independent of each other, the optical deformation scattering rate for a carrier with an initial wave vector of magnitude k_i into all final states of a given subband is finally given by:

$$W_{if}(k_i) = -\frac{D_o^2 m^*}{\rho\omega_o(2\pi)^2\hbar^2}\left(N_0 + \frac{1}{2} \mp \frac{1}{2}\right)$$

$$\times \int_0^\infty (G_{if}(K_z))^2\, \mathrm{d}K_z \int_0^{2\pi} \frac{k_i\cos\phi}{\sqrt{k_i^2\cos^2\phi - \frac{2m^*\Delta}{\hbar^2}}}\, \mathrm{d}\phi \quad (10.155)$$

10.14 Confined and interface phonon modes

Forming quantum wells or superlattices clearly changes the electronic energy levels of a crystal from what they are in an infinite bulk crystal, which is the main subject of this book. In

fact, all of the crystal properties are changed to a greater or lesser extent. Perhaps of secondary importance to the effect on the electronic energy levels, and the subsequent changes that this induces in scattering rates, exciton energies, impurity energies, etc. are the fundamental changes introduced to the *phonon modes*. It can be appreciated that such changes are likely as the bulk LO phonon in GaAs has an energy of around $36\,\mathrm{meV}$ while in AlAs it is closer to $50\,\mathrm{meV}$. Hence, forming a superlattice with alternating layers of GaAs and AlAs is going to have some effect on the phonon energies.

Various models have been put forward to account for this change in symmetry. At one end of the scale, some models consider each semiconductor layer as a continuum of material with macroscopic-like properties, namely the *dielectric continuum model* (see, for example, [9]) or the *hydrodynamic model* (see, for example, [15]). Alternative approaches have considered the allowed vibrational modes calculated directly from the viewpoint of individual atomic potentials (see, for example, [16]). See Adachi [10], p. 70, for an introduction.

Such improved models for phonons in heterostructures lead to modes which are confined to the individual semiconductor layers (*confined modes*) while some propagate along boundaries between the layers (the so-called *interface modes*). Recent work has shown that, while the electron–phonon scattering rates from these individual modes are quite different, the *total* rate from all of the modes collectively is quite similar to that from bulk phonons (see Kinsler *et al*. [17]). However, this is still a very active area of research, and future developments will need to be monitored.

10.15 Carrier–carrier scattering

Fermi's golden rule describes the lifetime of a particle in a particular state with respect to scattering by a *time-varying potential*. For phonon scattering, this harmonic potential is derived from the phonon wave function, which is itself a travelling wave. For the case of one carrier scattering against another due to the Coulomb potential, there appears to be no time dependency. The *Born approximation* is often cited in the literature when discussing carrier–carrier scattering; this is just a way of working scattering from a constant potential into Fermi's golden rule. This is achieved by considering that the perturbing potential is 'switched on' only when the particle reaches the same proximity. For an excellent introduction to the Born approximation, see Liboff [18], Chapter 14.

Therefore, the perturbing potential appearing in Fermi's golden rule for the interaction of two isolated carriers is the Coulombic interaction, i.e.

$$\tilde{\mathcal{H}} = \frac{e^2}{4\pi\epsilon r} \tag{10.156}$$

where $\epsilon = \epsilon_\mathrm{r}\epsilon_0$ is the permittivity of the material and r is the separation of the electrons. Now the initial and final states, $|i\rangle$ and $|f\rangle$ respectively, of the *system* both consist of two electron (or hole) wave functions, as carrier–carrier scattering is a two-body problem, and thus there is a much greater variety of scattering mechanisms possible than in the essentially one-body problem encountered in phonon scattering. Figure 10.17 illustrates all of the possible mechanisms in a two-level system, in which at least one of the carriers changes its subband; these are usually referred to as *intersubband* transitions. However, the distinction now is not quite so clear (see below). The central diagram in Fig. 10.17 illustrates the symmetric

intersubband event, $22 \rightarrow 11$, which moves two carriers down a level. The left and right figures show *Auger-type* intersubband scattering, where one carrier relaxes down to a lower subband, giving its excess energy to another carrier which remains within its original subband

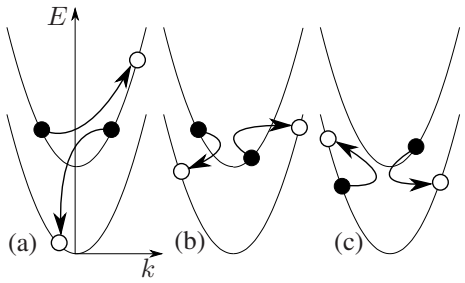


Figure 10.17: Illustration of intersubband carrier–carrier scattering mechanisms in a two-level system. The processes shown are (a) $22 \rightarrow 21$, (b) $22 \rightarrow 11$ and (c) $21 \rightarrow 11$

In addition to the above, there are also scattering events where the number of carriers in each subband does not change. Some of these are illustrated schematically in Fig. 10.18. Clearly the first of these, $22 \rightarrow 22$, is an intrasubband event; however, the second and third events are more difficult to categorise precisely because, although the number of carriers in each subband remains the same, the interaction itself is between carriers in different subbands. Pauli exclusion prevents carriers with the same spin occupying the same region of space, which therefore lowers their probability of scattering; in this work attention will be focused on collisions between particles with anti-parallel spins. Such considerations of spin-dependent scattering are often referred to by the term *exchange* [19]. Given that there are four possible carrier states involved, then in an N-level system there are 4^N different scattering events. In this two-level system these are as follows: 11–11, 11–12, 11–21, 11–22, 12–11, 12–12, 12–21, 12–22, 21–11, 21–12, 21–21, 21–22, 22–11, 22–12, 22–21 and 22–22. Note that completely different events of the type 'ij–fj' are possible in quantum wells with three or more subbands, and interactions of this type have been shown to be important in optically pumped intersubband lasers [19].

Therefore, taking a heterostructure wave function, of the form shown in equation (10.42), the matrix element in Fermi's golden rule (equation (10.16)) becomes:

$$\langle f|\tilde{\mathcal{H}}|i\rangle = \left\langle \psi_f(z)\frac{\mathrm{e}^{-i\mathbf{k}_f\bullet\mathbf{r}_{xy}}}{\sqrt{A}}\psi_g(z')\frac{\mathrm{e}^{-i\mathbf{k}_g\bullet\mathbf{r}'_{xy}}}{\sqrt{A}}\left|\frac{e^2}{4\pi\epsilon r}\right|\psi_i(z)\frac{\mathrm{e}^{-i\mathbf{k}_i\bullet\mathbf{r}_{xy}}}{\sqrt{A}}\psi_j(z')\frac{\mathrm{e}^{-i\mathbf{k}_j\bullet\mathbf{r}'_{xy}}}{\sqrt{A}}\right\rangle \tag{10.157}$$

where the subband indices of the initial states are labelled i and j, and those of the final states f and g. The decoupled form of the wave functions, with a component of the motion confined along the z-axis and an in-plane (x–y) travelling wave, suggests that the integrals should be evaluated across the plane and along the growth axis, and that the separation of the carriers be expressed as:

$$r = \sqrt{\left|\mathbf{r}_{xy} - \mathbf{r}'_{xy}\right|^2 + (z - z')^2} \tag{10.158}$$

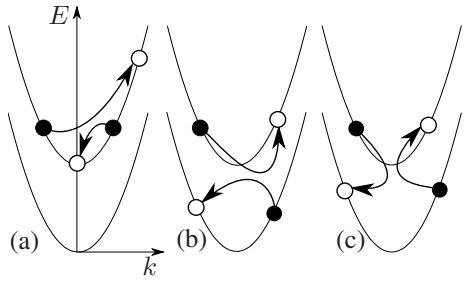


Figure 10.18: Illustration of various intrasubband carrier–carrier scattering mechanisms in a two-level system. The processes shown are (a) $22 \rightarrow 22$, (b) $21 \rightarrow 21$ and (c) $21 \rightarrow 12$

The matrix element is then expanded as:

$$\langle f|\tilde{\mathcal{H}}|i\rangle = \frac{e^2}{4\pi\epsilon A^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\iint \psi_f^*(z)\psi_g^*(z')\psi_i(z)\psi_j(z') \tag{10.159}$$
$$\times \frac{\mathrm{e}^{-i\left(\mathbf{k}_i\bullet\mathbf{r}_{xy}+\mathbf{k}_j\bullet\mathbf{r}'_{xy}\right)}\mathrm{e}^{i\left(\mathbf{k}_f\bullet\mathbf{r}_{xy}+\mathbf{k}_g\bullet\mathbf{r}'_{xy}\right)}}{\sqrt{\left|\mathbf{r}_{xy}-\mathbf{r}'_{xy}\right|^2+(z-z')^2}}\,\mathrm{d}\mathbf{r}'_{xy}\,\mathrm{d}\mathbf{r}_{xy}\,\mathrm{d}z'\,\mathrm{d}z$$

The scattering rate can now be derived by substituting this matrix element into Fermi's golden rule, as described by Goodnick and Lugli [20] (later reiterated by Smet *et al*. [21]), who followed earlier methods for carrier–carrier scattering in bulk (see, for example, Ziman [22], p. 170 and Takenaka *et al*. [23]).

Elaborating on this approach, first simplify the scattering matrix element by defining $\mathbf{q}_{xy} = \mathbf{k}_i - \mathbf{k}_f$ and $\mathbf{q}'_{xy} = \mathbf{k}_j - \mathbf{k}_g$, which represent the change in wave vector of the first and second carriers, respectively. A further substitution, $\mathbf{u} = \mathbf{r}_{xy} - \mathbf{r}'_{xy}$, is also used to represent the in-plane spatial separation between the two carriers. The matrix element therefore becomes:[3]

$$\langle f|\tilde{\mathcal{H}}|i\rangle = \frac{e^2}{4\pi\epsilon A^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\psi_f^*(z)\psi_g^*(z')\psi_i(z)\psi_j(z')\;\mathrm{d}z'\,\mathrm{d}z \tag{10.160}$$
$$\times \int \mathrm{e}^{-i(\mathbf{q}_{xy}+\mathbf{q}'_{xy})\bullet\mathbf{r}'_{xy}}\,\mathrm{d}\mathbf{r}'_{xy}\int\frac{\mathrm{e}^{-i\mathbf{q}_{xy}\bullet\mathbf{u}}}{\sqrt{u^2+(z-z')^2}}\,\mathrm{d}\mathbf{u}$$

The final integrals over $\mathbf{r}'_{xy}$ and $\mathbf{u}$ are both two-dimensional Fourier transforms and can be solved analytically using the standard results:

$$\int \mathrm{e}^{-i\mathbf{a}\bullet\mathbf{r}}\,\mathrm{d}\mathbf{r} = (2\pi)^2\delta(\mathbf{a}) \tag{10.161}$$

and

$$\int\frac{\mathrm{e}^{-i\mathbf{a}\bullet\mathbf{r}}}{\sqrt{r^2+b^2}}\,\mathrm{d}\mathbf{r} = \frac{2\pi}{a}\mathrm{e}^{-a|b|} \tag{10.162}$$

[3]Thanks to Emmanuel Dupont for useful discussions about improving this part of the derivation in the fourth edition.

Substituting these into the matrix element and expanding the Dirac δ-function in terms of the individual wave vectors yields:

$$\langle f|\tilde{\mathcal{H}}|i\rangle = \frac{(2\pi)^3 e^2}{4\pi\epsilon A^2 q_{xy}} A_{ijfg}(q_{xy})\ \delta(\mathbf{k}_f + \mathbf{k}_g - \mathbf{k}_i - \mathbf{k}_j) \tag{10.163}$$

where A_{ijfg} is a form factor, i.e.

$$A_{ijfg} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \psi_i(z)\psi_j(z')\psi_f^*(z)\psi_g^*(z')\ \mathrm{e}^{-q_{xy}|z-z'|}\ \mathrm{d}z'\ \mathrm{d}z \tag{10.164}$$

The Dirac δ-function in the matrix element enforces momentum conservation. It is interesting to note that this is an intrinsic property of carrier–carrier scattering, obtained directly from the derivation, and is not a 'manually' imposed restriction. Using the form for the matrix element shown in equation (10.163) and then substituting directly into Fermi's golden rule (equation (10.16)) gives the scattering rate for a pair of carriers initially in subbands i and j as follows:

$$W_{ijfg}(k_i, \mathbf{k}_j) = \frac{2\pi}{\hbar}\sum_{f,g}\left|\frac{(2\pi)^3 e^2}{4\pi\epsilon A^2 q_{xy}} A_{ijfg}(q_{xy})\ \delta(\mathbf{k}_f + \mathbf{k}_g - \mathbf{k}_i - \mathbf{k}_j)\right|^2 \times\ \delta(E_f^{\mathrm{t}} + E_g^{\mathrm{t}} - E_i^{\mathrm{t}} - E_j^{\mathrm{t}}) \tag{10.165}$$

Converting the summations over both final-state wave vectors into integrals introduces a factor of $L/(2\pi)$ per dimension, thus giving a factor of $A^2/(2\pi)^4$ in total (where the general area $A = L^2$) (see Section 2.4). Therefore:

$$W_{ijfg}(k_i, \mathbf{k}_j) = \frac{2\pi}{\hbar}\frac{A^2}{(2\pi)^4}\iint\left|\frac{(2\pi)^3 e^2}{4\pi\epsilon A^2 q_{xy}} A_{ijfg}(q_{xy})\ \delta(\mathbf{k}_f + \mathbf{k}_g - \mathbf{k}_i - \mathbf{k}_j)\right|^2 \times\ \delta(E_f^{\mathrm{t}} + E_g^{\mathrm{t}} - E_i^{\mathrm{t}} - E_j^{\mathrm{t}})\ \mathrm{d}\mathbf{k}_g\ \mathrm{d}\mathbf{k}_f \tag{10.166}$$

and so:

$$W_{ijfg}(k_i, \mathbf{k}_j) = \frac{(2\pi)^3 e^4}{\hbar A^2 (4\pi\epsilon)^2}\iint\frac{|A_{ijfg}(q_{xy})|^2}{q_{xy}^2}\ [\delta(\mathbf{k}_f + \mathbf{k}_g - \mathbf{k}_i - \mathbf{k}_j)]^2 \times\ \delta(E_f^{\mathrm{t}} + E_g^{\mathrm{t}} - E_i^{\mathrm{t}} - E_j^{\mathrm{t}})\ \mathrm{d}\mathbf{k}_g\ \mathrm{d}\mathbf{k}_f \tag{10.167}$$

Before proceeding further, it is necessary carefully to consider the meaning of the squared Dirac δ-function within the integral. Ordinarily, a δ-function 'sifts' an integral (i.e. selects just one point from an integrand) such that in two dimensions $\int f(\mathbf{k})\delta(\mathbf{k} - \mathbf{a})\ \mathrm{d}\mathbf{k} = f(\mathbf{a})$. However, if the δ-function is squared, then the expression is expanded as follows:

$$\begin{aligned}\int f(\mathbf{k})[\delta(\mathbf{k} - \mathbf{a})]^2\ \mathrm{d}\mathbf{k} &= \int f(\mathbf{k})\delta(\mathbf{k} - \mathbf{a})\delta(\mathbf{k} - \mathbf{a})\ \mathrm{d}\mathbf{k} \\ &= f(\mathbf{a})\delta(\mathbf{a} - \mathbf{a}) \\ &= f(\mathbf{a})\delta(\mathbf{0}) = \infty\end{aligned} \tag{10.168}$$

i.e. one of the δ-functions sifts the other, and obtains its value at the origin, which is infinitely large. In other words, *any* integral containing a squared δ-function evaluates to infinity and hence, at first appearance, equation (10.167) incorrectly predicts an infinitely large scattering rate.

To proceed, recall the original assumption about two-dimensional systems, that carriers are free to move in plane within a large (but finite) area A with length L on each side. This assumption was imposed in order to ensure that the wave functions are square-integrable. As such, equation (10.161) is only an approximate result because the system does not have infinite spatial extent. Equation (10.161) can be rearranged and written in approximate form for a system of finite area:

$$\delta(\mathbf{a}) \approx \frac{1}{(2\pi)^2} \int_{-A/2}^{A/2} \mathrm{e}^{-i\mathbf{a}\bullet\mathbf{r}} \, \mathrm{d}\mathbf{r} \tag{10.169}$$

By substituting $\mathbf{a} = \mathbf{0}$, the exponential term within the integral reduces to unity. The integral can be then be solved analytically to give an approximation to the height of the narrow two-dimensional δ-function 'spike' at the origin:

$$\delta(\mathbf{0}) \approx \frac{A}{(2\pi)^2} \tag{10.170}$$

Now, for systems with finite area, equation (10.169) can be written:

$$\int f(\mathbf{k})[\delta(\mathbf{k}-\mathbf{a})]^2 \, \mathrm{d}\mathbf{k} = f(\mathbf{a})\delta(\mathbf{0}) \approx f(\mathbf{a})\frac{A}{(2\pi)^2} \tag{10.171}$$

which is equivalent to sifting an integral with a single δ-function and multiplying by $A/(2\pi)^2$. This leads to the result:

$$\int f(\mathbf{k})[\delta(\mathbf{k}-\mathbf{a})]^2 \, \mathrm{d}\mathbf{k} \approx \frac{A}{(2\pi)^2} \int f(\mathbf{k})\delta(\mathbf{k}-\mathbf{a}) \, \mathrm{d}\mathbf{k} \tag{10.172}$$

Now, this result can be used to simplify the squared δ-function in equation (10.167) to obtain:

$$W_{ijfg}(k_i, \mathbf{k}_j) = \frac{2\pi e^4}{\hbar A(4\pi\epsilon)^2} \iint \frac{\left|A_{ijfg}(q_{xy})\right|^2}{q_{xy}^2} \, \delta(\mathbf{k}_f + \mathbf{k}_g - \mathbf{k}_i - \mathbf{k}_j)$$
$$\times \; \delta(E_f^{\mathrm{t}} + E_g^{\mathrm{t}} - E_i^{\mathrm{t}} - E_j^{\mathrm{t}}) \, \mathrm{d}\mathbf{k}_g \, \mathrm{d}\mathbf{k}_f \tag{10.173}$$

The total rate is then obtained by summing over all of the initial states of the second carrier (given by $\mathbf{k}_j$) and introducing Fermi–Dirac distribution functions to account for state occupancy:

$$W_{ijfg}(k_i, \mathbf{k}_j) = \frac{2\pi e^4}{\hbar A(4\pi\epsilon)^2} \sum_j \iint \frac{\left|A_{ijfg}(q_{xy})\right|^2}{q_{xy}^2} f_j^{\mathrm{FD}}(\mathbf{k}_j)\left[1 - f_f^{\mathrm{FD}}(\mathbf{k}_f)\right]\left[1 - f_g^{\mathrm{FD}}(\mathbf{k}_g)\right]$$
$$\times \; \delta(\mathbf{k}_f + \mathbf{k}_g - \mathbf{k}_i - \mathbf{k}_j)\delta(E_f^{\mathrm{t}} + E_g^{\mathrm{t}} - E_i^{\mathrm{t}} - E_j^{\mathrm{t}}) \, \mathrm{d}\mathbf{k}_g \, \mathrm{d}\mathbf{k}_f \tag{10.174}$$

As before, this summation can be replaced by an integral with the pre-factor $A/(2\pi)^2$ to obtain:

$$W_{ijfg}(k_i) = \frac{e^4}{2\pi\hbar(4\pi\epsilon)^2} \iiint \frac{|A_{ijfg}(q_{xy})|^2}{q_{xy}^2} f_j^{\text{FD}}(\mathbf{k}_j)\left[1 - f_f^{\text{FD}}(\mathbf{k}_f)\right] \tag{10.175}$$
$$\times \quad \left[1 - f_g^{\text{FD}}(\mathbf{k}_g)\right] \delta(\mathbf{k}_f + \mathbf{k}_g - \mathbf{k}_i - \mathbf{k}_j)\, \delta(E_f^{\text{t}} + E_g^{\text{t}} - E_i^{\text{t}} - E_j^{\text{t}})\, \mathrm{d}\mathbf{k}_g\, \mathrm{d}\mathbf{k}_f\, \mathrm{d}\mathbf{k}_j$$

which is equation (49) in Smet *et al*. [21]. Following their notation, collect the distribution functions together and label them as $P_{j,f,g}(\mathbf{k}_j, \mathbf{k}_f, \mathbf{k}_g)$. The first δ-function summarises in-plane momentum conservation and limits the integral over $\mathbf{k}_g$ to a contribution when $\mathbf{k}_g = \mathbf{k}_i + \mathbf{k}_j - \mathbf{k}_f$. In addition, the total energy of each carrier state is found by adding the in-plane kinetic energy to the subband minimum, as in equations (10.59) and (10.60). Thus:

$$W_{ijfg}(k_i) = \frac{e^4}{2\pi\hbar(4\pi\epsilon)^2} \iint \frac{|A_{ijfg}(q_{xy})|^2}{q_{xy}^2} P_{j,f,g}(\mathbf{k}_j, \mathbf{k}_f, \mathbf{k}_g)$$
$$\times \quad \delta\left(E_f + \frac{\hbar^2\mathbf{k}_f^2}{2m^*} + E_g + \frac{\hbar^2\mathbf{k}_g^2}{2m^*} - E_i - \frac{\hbar^2\mathbf{k}_i^2}{2m^*} - E_j - \frac{\hbar^2\mathbf{k}_j^2}{2m^*}\right) \mathrm{d}\mathbf{k}_f\, \mathrm{d}\mathbf{k}_j \tag{10.176}$$

and therefore:

$$W_{ijfg}(k_i) = \frac{m^* e^4}{\pi\hbar^3(4\pi\epsilon)^2} \iint \frac{|A_{ijfg}(q_{xy})|^2}{q_{xy}^2} P_{j,f,g}(\mathbf{k}_j, \mathbf{k}_f, \mathbf{k}_g)$$
$$\times \quad \delta\left(\mathbf{k}_f^2 + \mathbf{k}_g^2 - \mathbf{k}_i^2 - \mathbf{k}_j^2 + \frac{2m^*}{\hbar^2}(E_f + E_g - E_i - E_j)\right) \mathrm{d}\mathbf{k}_f\, \mathrm{d}\mathbf{k}_j \tag{10.177}$$

where $\mathbf{k}_g$ is known in terms of the other three wave vectors. It is the assumption of parabolic subbands in this last step that will be the limiting factor for the application of this method to hole–hole scattering—a point mentioned earlier in the context of the carrier–LO phonon scattering rate derivation. Now equation (10.177) represents the scattering rate of a carrier at a particular wave vector $\mathbf{k}_i$ averaged over all of the other initial particle states $\mathbf{k}_j$, and hence the only unknown in this remaining δ-function is the wave vector $\mathbf{k}_f$. Contributions to the integral over $\mathbf{k}_f$ occur when the argument of the δ-function is zero, and indeed, given the form for this argument, it is clear that the solutions for $\mathbf{k}_f$ map out an ellipse.

The standard procedure [20, 21] is then to introduce relative wave vectors:

$$\mathbf{k}_{ij} = \mathbf{k}_j - \mathbf{k}_i \tag{10.178}$$
$$\mathbf{k}_{fg} = \mathbf{k}_g - \mathbf{k}_f \tag{10.179}$$

and replace the integration over $\mathbf{k}_f$ in equation (10.177) with an integration over $\mathbf{k}_{fg}$. Since:

$$\mathbf{k}_f = \mathbf{k}_i + \mathbf{k}_j - \mathbf{k}_g = \mathbf{k}_i + \mathbf{k}_j - \mathbf{k}_{fg} - \mathbf{k}_f \tag{10.180}$$

i.e.

$$\mathbf{k}_f = \frac{1}{2}(\mathbf{k}_i + \mathbf{k}_j - \mathbf{k}_{fg}) \tag{10.181}$$

it follows that:

$$\mathrm{d}(\mathbf{k}_f)_x = -\frac{1}{2}\mathrm{d}(\mathbf{k}_{fg})_x \quad \text{and} \quad \mathrm{d}(\mathbf{k}_f)_y = -\frac{1}{2}\mathrm{d}(\mathbf{k}_{fg})_y \tag{10.182}$$

so:

$$\mathrm{d}\mathbf{k}_f = \frac{1}{4}\mathrm{d}\mathbf{k}_{fg} \tag{10.183}$$

as pointed out by Moško [24], which should then be substituted into equation (10.177). In order to perform the integration over $\mathbf{k}_{fg}$ it is converted to plane polar coordinates with $\mathrm{d}\mathbf{k}_{fg} = k_{fg}\mathrm{d}k_{fg}\mathrm{d}\theta$, where θ is an angle measured from $\mathbf{k}_{ij}$ and the trajectory in the k_{fg}–θ plane is deduced from the condition that the argument of the δ-function in equation (10.177) must be zero.

The conservation of momentum diagram therefore looks like Fig. 10.19, where:

$$\mathbf{k}_{\text{sum}} = \mathbf{k}_i + \mathbf{k}_j = \mathbf{k}_f + \mathbf{k}_g \tag{10.184}$$

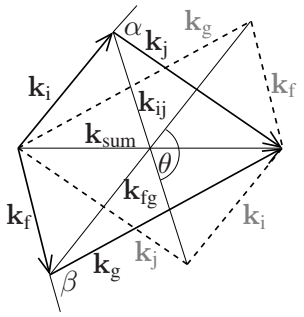


Figure 10.19: The conservation of momentum diagram for the two-body electron–electron scattering event

In this work, the occupancy of the final states will be assumed to be small, such that the distribution functions dependent upon $\mathbf{k}_f$ and $\mathbf{k}_g$ can be ignored. The effect of setting $f_f^{\mathrm{FD}}(\mathbf{k}_f)$ and $f_g^{\mathrm{FD}}(\mathbf{k}_g)$ to zero can be seen to place an upper limit on the scattering rate. The physical interpretation of this approximation is that *final-state blocking* is ignored, i.e. the process by which a scattering event is prevented because the required final state is already occupied. This is a common simplification [25] which allows the particular values of the final-state wave vectors to remain unknown. The approximation that this introduces is 'good' for the typical carrier densities encountered in devices based on quantum well heterostructures [26]. Recently, a generalisation of this approach has been put forward that allows the specific values of the final carrier wave vectors to be calculated, thus avoiding this approximation; this method is also an example of making a different choice for the angle θ [19].

Using the notation $k = |\mathbf{k}|$, and applying the cosine rule to Fig. 10.19:

$$k_{\text{sum}}^2 = k_i^2 + k_j^2 + 2k_i k_j \cos\alpha \tag{10.185}$$

$$k_{\text{sum}}^2 = k_f^2 + k_g^2 + 2k_f k_g \cos\beta \tag{10.186}$$

and also:

$$k_{ij}^2 = k_i^2 + k_j^2 - 2k_i k_j \cos\alpha \tag{10.187}$$

$$k_{fg}^2 = k_f^2 + k_g^2 - 2k_f k_g \cos\beta \tag{10.188}$$

Summing equation (10.185) with equation (10.187) and (10.186) with (10.188) gives:

$$k_{\text{sum}}^2 + k_{ij}^2 = 2k_i^2 + 2k_j^2 \tag{10.189}$$

$$k_{\text{sum}}^2 + k_{fg}^2 = 2k_f^2 + 2k_g^2 \tag{10.190}$$

and eliminating k_{sum}^2 gives:

$$k_{fg}^2 = k_{ij}^2 + 2\left(k_f^2 + k_g^2 - k_i^2 - k_j^2\right) \tag{10.191}$$

Energy conservation has already been considered earlier within the δ-function, i.e.

$$E_f + \frac{\hbar^2\mathbf{k}_f^2}{2m^*} + E_g + \frac{\hbar^2\mathbf{k}_g^2}{2m^*} = E_i + \frac{\hbar^2\mathbf{k}_i^2}{2m^*} + E_j + \frac{\hbar^2\mathbf{k}_j^2}{2m^*} \tag{10.192}$$

thus giving:

$$k_f^2 + k_g^2 - k_i^2 - k_j^2 = \frac{2m^*}{\hbar^2}\left(E_i + E_j - E_f - E_g\right) \tag{10.193}$$

Substituting into equation (10.191) then gives:

$$k_{fg}^2 = k_{ij}^2 + \frac{4m^*}{\hbar^2}\left(E_i + E_j - E_f - E_g\right) = k_{ij}^2 + \Delta k_0^2 \tag{10.194}$$

say, where the notation of Smet *et al*. has been employed.

Consider the entity $\mathbf{q}_{xy}$, for which q_{xy} is the magnitude. This vector $\mathbf{q}_{xy} = \mathbf{k}_i - \mathbf{k}_f$ is illustrated in the equivalent diagram (Fig. 10.20). In addition, this figure emphasises the difference between the relative wave vectors $\mathbf{k}_{ij}$ and $\mathbf{k}_{fg}$, which turns out to be equal to $2\mathbf{q}_{xy}$, and therefore:

$$\mathbf{q}_{xy} = \frac{\mathbf{k}_{ij} - \mathbf{k}_{fg}}{2} \tag{10.195}$$

With an angle of θ between them, the cosine rule then gives:

$$(2q_{xy})^2 = k_{ij}^2 + k_{fg}^2 - 2k_{ij}k_{fg}\cos\theta \tag{10.196}$$

As the aim is to eliminate the final-state wave vectors from the definition of q_{xy}, substitute for k_{fg} from equation (10.194), thus giving:

$$(2q_{xy})^2 = 2k_{ij}^2 + \Delta k_0^2 - 2k_{ij}\sqrt{k_{ij}^2 + \Delta k_0^2}\cos\theta \tag{10.197}$$

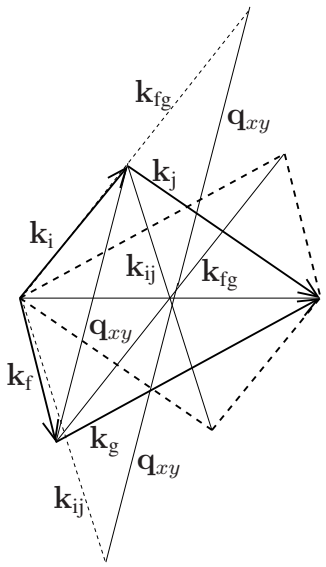


Figure 10.20: The conservation of momentum diagram for the two-body electron–electron scattering event

Summarising, then, the expression for the scattering rate of a carrier of a particular wave vector $\mathbf{k}_i$ with another carrier is given by:

$$W_{ijfg}(k_i) = \frac{m^* e^4}{4\pi\hbar^3(4\pi\epsilon)^2} \int\int_0^{2\pi} \frac{|A_{ijfg}(q_{xy})|^2}{q_{xy}^2} \, P_{j,f,g}(\mathbf{k}_j, \mathbf{k}_f, \mathbf{k}_g) \, \mathrm{d}\theta \, \mathrm{d}\mathbf{k}_j \qquad (10.198)$$

which is equation (51) in Smet *et al.* [21], but with an additional factor of 4 in the denominator [24]. In the present treatise the factor $P_{jfg}(\mathbf{k}_j, \mathbf{k}_f, \mathbf{k}_g)$ has been assumed to be dependent upon the initial-state distribution only, i.e. $P_{jfg}(\mathbf{k}_j, \mathbf{k}_f, \mathbf{k}_g) = f_j^{\mathrm{FD}}(\mathbf{k}_j)$, and q_{xy} is given by equation (10.197). The integration over the vector $\mathbf{k}_j$ can be performed by effectively switching to plane polar coordinates, and integrating along the length k_{ij} and around the angle α between $\mathbf{k}_i$ and $\mathbf{k}_j$.

With this change in the integral, equation (10.198) then becomes:

$$W_{ijfg}(k_i) = \frac{m^* e^4}{4\pi\hbar^3(4\pi\epsilon)^2} \int\int_0^{2\pi}\int_0^{2\pi} \frac{|A_{ijfg}(q_{xy})|^2}{q_{xy}^2} \, P_j(\mathbf{k}_j) \, \mathrm{d}\theta \, \mathrm{d}\alpha \, k_j \, \mathrm{d}k_j \qquad (10.199)$$

where q_{xy} is given by equation (10.197), Δk_0^2 is given by equation (10.194), and from Fig. 10.21:

$$k_{ij}^2 = k_i^2 + k_j^2 - 2k_i k_j \cos\alpha \qquad (10.200)$$

As in the case of phonon scattering, the limits of integration of the wave vector are taken from zero (the subband minima) to that corresponding to the barrier height of the heterostructure in question.

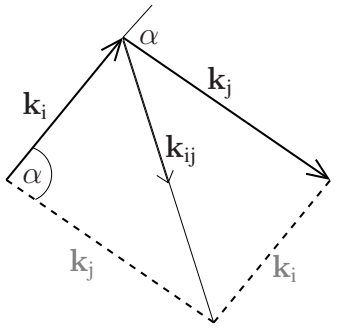


Figure 10.21: The relationship between the two initial carrier states

10.16 Addition of screening

So far, the theoretical treatise has only considered carrier–carrier scattering of two carriers in an empty environment. In semiconductor heterostructures, carrier–carrier scattering occurs because the system is doped, and therefore by definition there will be many carriers. In such instances, the force between any two carriers is not just the bare Coulombic repulsion, as the other mobile carriers are able to respond to any change in the electrostatic field, with the result being a reduction in the probability of scattering; the other carriers are said to *screen* the interaction.

One of the simplest models for screening [21] considers only the carriers within the same subband as the initial carrier state; it then proceeds by replacing the dielectric constant $\epsilon = \epsilon_{\mathrm{r}}\epsilon_0$ with one which is dependent upon the relative wave vector q_{xy}, i.e. $\epsilon = \epsilon_{\mathrm{r}}\epsilon_0\epsilon_{\mathrm{sc}}$, where:

$$\epsilon_{\mathrm{sc}} = 1 + \frac{2\pi e^2}{(4\pi\epsilon)q_{xy}}\Pi_{ii}(q_{xy}, E_{\mathrm{F}}, T)A_{iiii}(q_{xy}) \tag{10.201}$$

One consequence of this replacement is that the $q_{xy} = 0$ pole is removed from the scattering rate. The polarisation factor $\Pi_{ii}(q_{xy}, E_{\mathrm{F}}, T)$ is taken from Maldague [27], but the dependency on angular frequency ω has been dropped because only the static ($w = 0$) case is being considered. Note also that the double index ii implies that only carriers within a single (the initial) subband are considered, hence the approach is known as *static single-subband screening*. Combining the results of Ando *et al*. [28] and using a Heaviside unit step function for convenience, the polarisation factor at absolute zero is then given by:

$$\Pi_{ii}(q_{xy}, E_{\mathrm{F}}, T = 0) = \frac{m^*}{\pi\hbar^2}\left[1 - \Theta(q_{xy} - 2k_{\mathrm{F}})\sqrt{1 - \left(\frac{2k_{\mathrm{F}}}{q_{xy}}\right)^2}\right] \tag{10.202}$$

where k_{F} is the two-dimensional Fermi wave vector [28] applicable to the initial state i and is defined only at absolute zero, i.e.

$$k_{\mathrm{F}} = \sqrt{\frac{2\pi N_i}{g}} \tag{10.203}$$

(see [6], p. 36, for example, for the three-dimensional equivalent). The factor g accounts for any degeneracy which can lead to multiple subband valleys; in GaAs, this is just 1. This result can also be derived from equation (2.85) by taking the limit of low T.

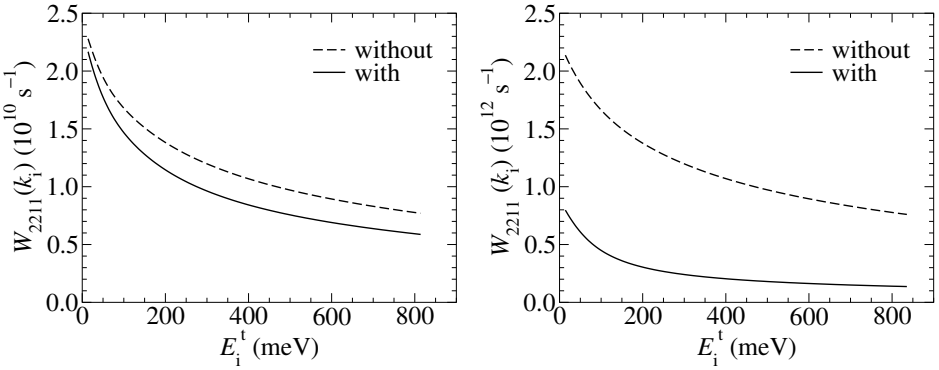


Figure 10.22: The electron–electron scattering rate as a function of the initial electron energy, with and without screening, for a carrier density of $1 \times 10^{10}\,\mathrm{cm}^{-2}$ (left) and $100 \times 10^{10}\,\mathrm{cm}^{-2}$ (right)

Maldague [27] and Smet [21] showed that, under certain approximations, this polarisability can be generalised to any temperature with:

$$\Pi_{ii}(q_{xy}, E_\mathrm{F}, T) = \int_0^\infty \frac{\Pi_{ii}(q_{xy}, E, T=0)}{4kT\cosh^2\left((E_\mathrm{F}-E)/(2kT)\right)}\,\mathrm{d}E \tag{10.204}$$

where E_F is the quasi-Fermi energy of the initial state i and the integration is performed over the subband energies of state i. Note that the numerator $\Pi_{ii}(q_{xy}, E, T=0)$ in the integrand *looks like* the zero temperature polarisability, except that the argument E is the variable of integration E,[4] i.e.

$$\Pi_{ii}(q_{xy}, E, T=0) = \frac{m^*}{\pi\hbar^2}\left[1 - \Theta(q_{xy}-2k)\sqrt{1-\left(\frac{2k}{q_{xy}}\right)^2}\right] \tag{10.205}$$

where $E = \hbar^2k^2/(2m^*)$.

Figure 10.22 shows the results of a calculation of the 22–11 electron–electron scattering rate as a function of the electron energy of the initial state i, in an infinitely deep $400\,\text{Å}$ wide GaAs quantum well. The figure compares the rates with and without the screening term at a temperature of 77 K and for carrier densities of $1 \times 10^{10}\,\mathrm{cm}^{-2}$ (left) and $100 \times 10^{10}\,\mathrm{cm}^{-2}$ (right) in each level.

The first conclusion to be drawn is that the scattering rate increases as the energy of the initial electron in state i decreases towards the subband minimum. In addition, the scattering

[4] Not the constant of the quasi-Fermi energy as in earlier editions, thanks to Chris Baird [29, 30].

rates themselves are nearly two orders of magnitude higher in the higher-carrier-density case illustrated in Fig. 10.22 (right) than in the lower-density case (left). Furthermore, it is apparent that screening reduces the scattering rate, and the higher the carrier density, the larger the effect of screening, as might be expected.

10.17 Mean intersubband carrier–carrier scattering rate

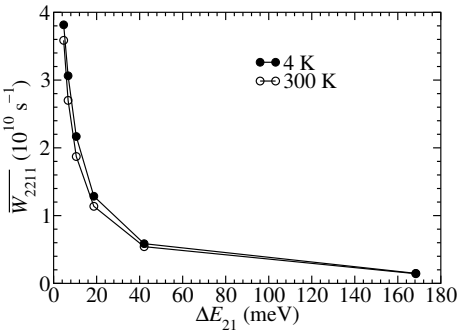


Figure 10.23: The mean 22–11 electron–electron scattering rate as a function of the subband separation, including screening and with a carrier density of $1 \times 10^{10}\,\mathrm{cm}^{-2}$ in each level

The method documented so far gives the carrier–carrier scattering rate for a particular carrier energy averaged over another initial state distribution; as before for the carrier–LO phonon case, it is usually more instructive to average this quantity over the Fermi–Dirac distribution of carriers in the initial state, as described in Section 10.4.

Figure 10.23 shows the electron–electron scattering rate averaged over the initial carrier distribution, as given by equation (10.28), for the 22–11 mechanism in an infinitely deep quantum well. The subband separation, $\Delta E_{21} = E_2 - E_1$, was varied by adjusting the quantum well width. It can be seen that the electron–electron scattering rate increases as the subband separation decreases and that in this particular approach, with the low carrier density of $1 \times 10^{10}\,\mathrm{cm}^{-2}$, the rate is only a weak function of temperature.

The effect of increasing carrier density is shown in Fig. 10.24(a). The range of carrier densities was chosen to span roughly those found in intersubband devices. It can be seen that, with this model, the intersubband scattering is almost proportional to the carrier density, although it must be noted that in this work any effects due to Pauli exclusion are being ignored—this took the form of assuming that the final-state populations were small, as discussed earlier. Inclusion of significant populations in a final-state could lead to final-state blocking being significant, and hence interfere with the proportionality, but at this stage of gaining an intuitive feel for carrier–carrier scattering, such investigations necessitate the introduction of yet more variables and unnecessary complication. As in the previous example, it was found that these intersubband scattering rates were weakly sensitive to the temperature. This is substantiated further in Fig. 10.24(b), which shows the temperature dependence of the

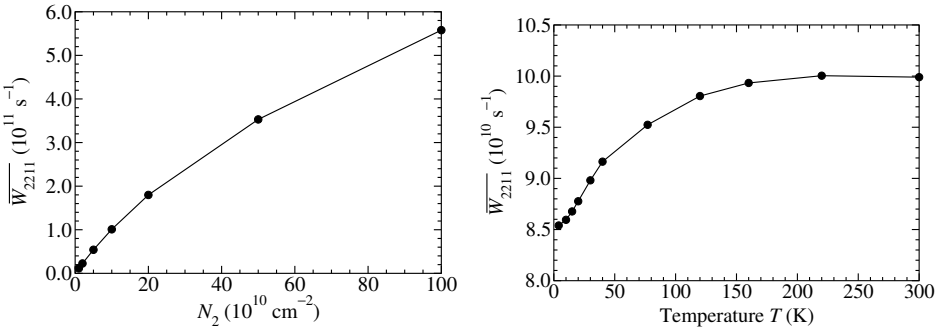


Figure 10.24: The mean 22–11 electron–electron scattering rate for a 300 Å infinitely deep GaAs quantum well (a) as a function of the carrier density in level $|2\rangle$ at a fixed temperature of 300 K and (b) as a function of the temperature, using a fixed carrier density of $10 \times 10^{10}\,\mathrm{cm}^{-2}$

22–11 intersubband scattering rate for $10 \times 10^{10}\,\mathrm{cm}^{-2}$ electrons in the same 300 Å infinitely deep quantum well. The rate increases by around 15% as the temperature increases towards 300 K. The temperature dependency within the model comes from the screening term as well as both of the initial-state distribution functions.

10.18 Computational implementation

At this stage, it is worth considering the computation methods used to solve equation (10.199). The scattering matrix element $A_{ijfg}(q_{xy})$ (equation (10.164)) is an integral over two variables (i.e. the locations of each carrier). When solved numerically (e.g. using the trapezium rule), the integrand in the equation is sampled at a number of discrete spatial points, N_z. When integrated over both z and z', the total number of samples is therefore N_z^2. This matrix element must, in turn, be integrated over three variables in reciprocal space (the angles α and θ and the wave vector k_j), each of which would require a number of samples, N_k. As such, a direct evaluation of the numerical integral in each dimension can become extremely time-consuming. For example, taking a modest number of steps ($N_z = 1000$, $N_k = 100$) in each dimension would require the integrand of $A_{ijfg}(q_{xy})$ to be computed $N_z^2 N_k^3 = 1 \times 10^{12}$ times! Clearly, this is a formidable computation task, and it is desirable to cast the calculation into a more computationally efficient form.

First, it is important to note that the scattering matrix element does not depend *directly* upon α, θ or k_j, and instead is only a function of q_{xy}. As such, it is inefficient to recalculate the matrix element for each combination of the three variables. It is helpful to rewrite the integrand in equation (10.199) in a simplified form:

$$W_{ijfg}(k_i) = \frac{m^* e^4}{4\pi\hbar^3(4\pi\epsilon)^2} \int P_j(\mathbf{k}_j) k_j \int_0^{2\pi} \int_0^{2\pi} F_{ijfg}(q_{xy}) \;\mathrm{d}\theta\,\mathrm{d}\alpha\,\mathrm{d}k_j \qquad (10.206)$$

where the new matrix element is given by:

$$F_{ijfg}(q_{xy}) = \frac{|A_{ijfg}(q_{xy})|^2}{(q_{xy}\epsilon_{\text{sc}})^2} \tag{10.207}$$

Note that the $P_j(\mathbf{k}_j)k_j$ term has been taken outside the integrals over angle. The main advantage of this form of the equation is that the new matrix element $F_{ijfg}(q_{xy})$ can be tabulated in a look-up table for a range of values of q_{xy} *before commencing* the three-dimensional integration in equation (10.206). The appropriate value of the matrix element can then simply be found in the table, much faster than recalculating it at each step within the integration.

A second significant improvement can be made by optimising the calculation of $A_{ijfg}(q_{xy})$ itself. For this purpose, it is helpful to rewrite the matrix element in the form:

$$A_{ijfg}(q_{xy}) = \int \psi_i(z)\psi_f^*(z)I_{jg}(q_{xy}, z)\,\mathrm{d}z \tag{10.208}$$

where the integral over z' has been absorbed into a new matrix element:

$$I_{jg}(q_{xy}, z) = \int \psi_j(z')\psi_g^*(z')\mathrm{e}^{-q_{xy}|z-z'|}\,\mathrm{d}z' \tag{10.209}$$

The calculation of this matrix element constitutes the innermost integral, and it must therefore be evaluated many times. Even seemingly minor reductions in the time required for this step have a significant effect on the total computation effort. To this end, the integrand may be split into a pair of functions with semi-infinite domains:

$$\mathrm{e}^{-q_{xy}|z-z'|} = \begin{cases} \mathrm{e}^{q_{xy}(z'-z)} = \mathrm{e}^{q_{xy}z'}/\mathrm{e}^{q_{xy}z}, & z \geq z' \\ \mathrm{e}^{q_{xy}(z-z')} = \mathrm{e}^{q_{xy}z}/\mathrm{e}^{q_{xy}z'}, & z < z' \end{cases} \tag{10.210}$$

It follows that the matrix element may be rewritten in the form:

$$I_{jg}(q_{xy}, z) = \frac{1}{\mathrm{e}^{q_{xy}z}}C_{jg}^{-}(q_{xy}, z) + \mathrm{e}^{q_{xy}z}C_{jg}^{+}(q_{xy}, z) \tag{10.211}$$

where:

$$C_{jg}^{-}(q_{xy}, z) = \int_{-\infty}^{z} \psi_g^*(z')\mathrm{e}^{q_\alpha z'}\psi_j(z')\,\mathrm{d}z' \tag{10.212}$$

$$C_{jg}^{+}(q_{xy}, z) = \int_{z}^{\infty} \psi_g^*(z')\frac{1}{\mathrm{e}^{q_{xy}z'}}\psi_j(z')\,\mathrm{d}z'$$

Although this expression appears to be more complex *mathematically*, since a single integral has been replaced by a pair, it is in fact simpler *computationally*. This is because the modulus function no longer appears in the expression, and as such, the values of the exponential function $\exp(q_{xy}z)$ can be tabulated and reused in both parts of equation (10.212). Although software optimisation is not the focus of this present work, it is important to note from this discussion that the most efficient computational solution to a problem is not always the most elegant or intuitively mathematical expression. As such, a dilemma is often encountered in computational physics between writing the 'neatest' and the fastest code.

10.19 Intrasubband versus intersubband

So far, only the results of calculations of intersubband carrier–carrier scattering rates have been presented, but as mentioned before, intrasubband events are also possible. The latter are characterised by no change in the numbers of carriers in each subband. Considering only the most simplistic mechanisms of the type ii–ii, Fig. 10.25(a) plots the carrier density dependency and Fig. 10.25(b) the temperature dependency of the 22–22 electron–electron scattering rate, again for a typical 300 Å infinitely deep GaAs quantum well. These figures are the intrasubband equivalents of the intersubband cases of Fig. 10.24(a) and (b), respectively. Note that, as all of the initial and final states are within the same subband, no dependence on an intersubband energy separation is required.

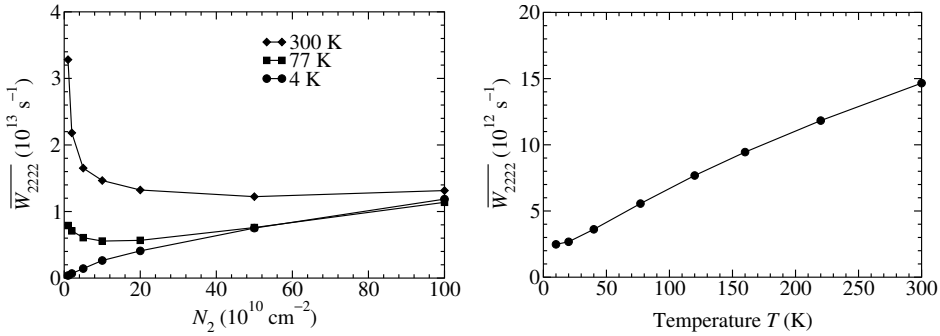


Figure 10.25: The mean 22–22 electron–electron scattering rate for a 300 Å infinitely deep GaAs quantum well (a) as a function of the carrier density at a fixed temperature of 300 K and (b) as a function of the temperature, using a fixed carrier density of $10 \times 10^{10}\,\mathrm{cm}^{-2}$

Therefore, comparing Fig. 10.25(a) with Fig. 10.24(a), it can be seen that the dependence of the scattering rate on the carrier density is quite different in the intrasubband case. The 'almost' proportionality is evident at 4 K, but for 77 K and 300 K there is a clear non-monotonic behaviour.

Figure 10.25(b) shows the temperature dependence of this 22–22 rate; this is stronger than in the intersubband case of Fig. 10.24(b), increasing approximately linearly between 4 K and 300 K.

Overall, the intrasubband scattering rates are around one to two orders of magnitude higher than the closest intersubband equivalent. This is a general result which is observed in a range of semiconductor heterostructures.

10.20 Thermalised distributions

This last result, i.e. that the intrasubband scattering rates are much larger than their typical intersubband equivalents, is very important. Its implication is that, given a quantum well or a quantum wire (basically any system which has carrier dispersion) with a number of

subbands, the carriers within each subband scatter much more rapidly than they do between the subbands. Thus, if the system is stimulated somehow, either by optical excitation or electrical injection of carriers, then the distributions of carriers within each subband will, in the first instance, reorganise themselves independently of each other, before intersubband scattering begins to redistribute the carriers between the levels. This justifies the initial assumption that the carrier populations in each subband can be represented by Fermi–Dirac distributions, although perhaps with independent quasi-Fermi energies (see Section 2.5)—in this situation, the carriers are said to be *independently thermalised*. This is a common assumption in intersubband devices (see, for example, [31]). It is also possible that the temperature required to describe these distributions is not equal to the lattice temperature; such a scenario will either require a thorough analysis of the kinetics of the system or else a Monte Carlo simulation of the complete subband structure.

Should the means of excitation, i.e. the source of input energy, be removed from the quantum well system, then the carriers will relax down to a state where the populations can all be described by a single Fermi energy—this would represent thermal equilibrium. Thus, the previous case of an excited subband structure is often given the rather grand descriptive title of a *non-equilibrium carrier distribution*.

10.21 Auger-type intersubband processes

The various carrier–carrier scattering rates were discussed earlier, and illustrated in Figs. 10.17 and 10.18. Until now, the calculations themselves have only considered the 'symmetric' 22–11 process, and only within an infinitely deep quantum well. This particular heterostructure was just an illustrative example, and provided a way of reducing the number of material parameters that had to be specified. In order to calculate the scattering rates due to the 'asymmetric' 22–21 and 21–11 processes, it is necessary to base the calculations in a different quantum well system; for this purpose, a single 200 Å GaAs quantum well surrounded by $Ga_{0.8}Al_{0.2}As$ barriers was chosen. This is also an opportunity to reiterate and demonstrate that the carrier–carrier scattering rate derived is valid for any one-particle eigenstates, as calculated with the methods outlined in Chapters 2 and 3.

The asymmetric intersubband processes are often referred to as *Auger-type processes* because they resemble traditional interband Auger scattering, where one carrier is able to give up its potential energy to another carrier and hence relax down a level. These processes are forbidden in symmetric potentials, and the simplest way to break the symmetry of the above finite quantum well is to apply an electric field. Figure 10.26 illustrates the potential profile and the wave functions of the ground $|1\rangle$ and first excited $|2\rangle$ states of the quantum well with the maximum applied electric field employed.

Fig. 10.27 shows the results of calculations of the mean scattering rates for all three possible intersubband mechanisms in this two-level system. The selection rules arise quite naturally in the formalism; hence, at zero field when the potential and hence the wave functions are symmetric (and anti-symmetric), only the symmetric 22–11 channel is allowed. However, as the electric field is increased and the symmetry of the potential broken, the asymmetric rates increase and the symmetric rate decreases. At the maximum value of the electric field employed, the Auger-type rates are a significant fraction of the total scattering rate, thus illustrating that their inclusion is essential in modelling such systems. Indeed,

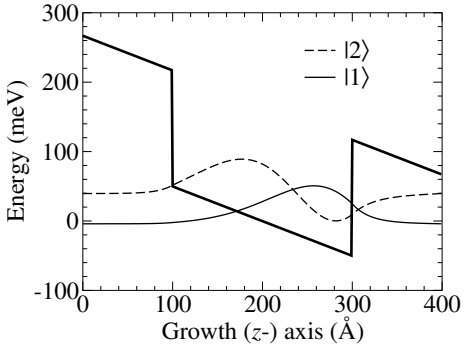


Figure 10.26: The potential profile and two lowest-energy eigenstates of a $200\,\text{Å}$ GaAs quantum well surrounded by $Ga_{0.8}Al_{0.2}As$ barriers, at an applied electric field of $50\,\text{kV}\,\text{cm}^{-1}$

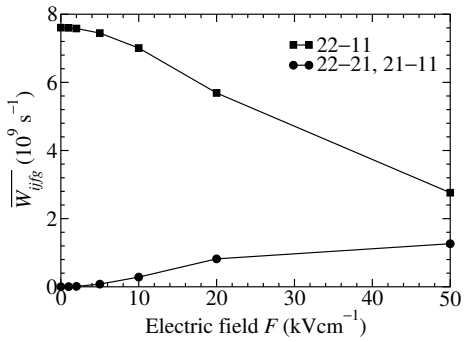


Figure 10.27: The mean intersubband electron–electron scattering rate as a function of the applied electric field for the quantum well in Fig. 10.26 with 1×10^{10} electrons per square centimetre in each level; note that the scattering rates for the 22–21 and 21–11 processes are almost identical

quantum well systems exist in which the Auger-type scattering rates dominate (see, for example, [32]). Note that the 22–11 rate moves two electrons down a level, whereas the two Auger-type rates only move one electron down at a time; this fact must be accounted for when solving subband population rate equations.

10.22 Asymmetric intrasubband processes

For completeness, Fig. 10.28 shows the scattering rates of possible intrasubband mechanisms for the same finite quantum well as in the previous section. With equal carrier densities employed in each level, the 'asymmetric' 21–21 and 21–12 intrasubband events are significant too; particularly the former. The implication of these results is that there will be a

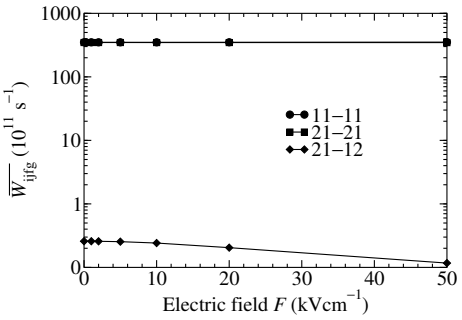


Figure 10.28: The mean intrasubband electron–electron scattering rates as a function of the applied electric field for the quantum well in Fig. 10.26 with a carrier density of $1 \times 10^{10}\,\mathrm{cm}^{-2}$ in each level. Note that the 21–21 and 11–11 rates are almost identical

substantial interaction between the carrier distributions in the different subbands. Therefore, even disregarding the transfer of carriers between levels with intersubband scattering, the distributions are coupled in some way and will not be independently thermalised, as often conjectured. The high scattering rate of the 21–21 channel implies that there will be a redistribution of kinetic energy between the subbands without a redistribution of carriers, and thus it seems likely that, although the electron temperature will not be equal to that of the lattice, the same electron temperature may be applicable to the entire multi-subband carrier distribution.

10.23 Empirical relationships

Simple empirical relationships giving the dependency of some computationally complex quantity, such as a carrier scattering rate, on some simple material parameter, are often useful for device designers and experimentalists who need an intuitive feel for the underlying physics without having recourse to lengthy calculations. Earlier, a relationship was derived for the ratio of emission to absorption rates for carrier–LO phonon scattering (see equation (10.113)). As hinted at then, the subband separation and temperature dependencies of the 22–11 carrier–carrier scattering rate imply simple relationships (see Section 10.17). In particular, the hyperbolic energy dependence suggests that:

$$\overline{W_{2211}} \propto \frac{1}{\Delta E_{21}} \tag{10.213}$$

and the linear carrier density dependence:

$$\overline{W_{2211}} \propto N \tag{10.214}$$

where, in this simple example of a symmetric ii–ff rate, the N refers to the carrier density in the initial state i.

The diagrams shown in Fig. 10.29 substantiate these postulated relationships, i.e. the graphs are quite linear and close to the origin. Thus, given a quantum well system with

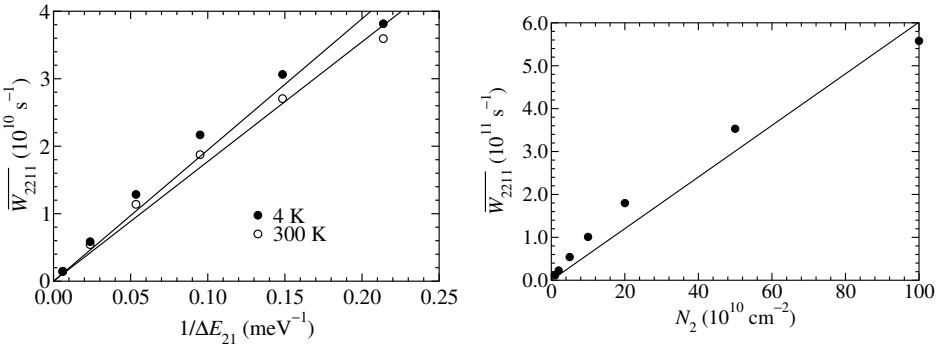


Figure 10.29: The mean 22–11 intersubband electron–electron scattering rate plotted against the reciprocal of the subband separation (left) and the carrier density (right)

a particular doping profile, and hence a specific carrier density, it is possible to answer questions such as 'what happens to the intersubband electron–electron scattering rate if we increase the well width and hence halve the subband separation?'. The answer is, of course, that the rate will double. Similar questions can be asked about variations in doping profile and hence carrier densities for fixed layer thicknesses.

Both results can be combined as follows:

$$\overline{W_{2211}} \propto \frac{N}{\Delta E_{21}} \tag{10.215}$$

10.24 A generalised expression for scattering of two-dimensional carriers

The previous sections have shown that Fermi's golden rule can be used directly to derive the scattering rates for a given perturbing potential within a two-dimensional system. At this stage, though, it is worth noting that these derivations shared many common steps. Indeed, it is possible to derive a more specialised form of Fermi's golden rule, for two-dimensional systems, which can be used as a starting point for the derivation of scattering rates.

Starting from the general expression for the scattering rate, equation (10.15), the *total* energy of a state within a subband can be written:

$$E_j^{\mathrm{t}} = E_j + \frac{\hbar^2 k_j^2}{2m^*} \tag{10.216}$$

where j is the index of the subband, E_j is the energy at the bottom of the subband and k_j is the in-plane wave vector. Substituting this expression for the total energies of the initial and final states into the energy conservation term in equation (10.15) gives:

$$W_{if}(k_i, k_\alpha) = \frac{2\pi}{\hbar} \left| \langle f | \tilde{\mathcal{H}} | i \rangle \right|^2 \delta \left[E_f - E_i + \frac{\hbar^2}{2m^*} \left(k_\alpha^2 - k_i^2 \right) \right] \Theta \left(k_\alpha^2 \right) \tag{10.217}$$

Here, k_α is the wave vector of an arbitrary state in the final subband. The Θ term at the end of this expression is the *Heaviside* step function, which ensures that the rate is non-zero only if the final state lies within the final subband (i.e. the wave vector must be real).

Now use can be made of the property of δ-functions that:

$$\delta[f(k)] = \sum_n \frac{\delta(k - k_n)}{|f'(k_n)|} \tag{10.218}$$

where $f(k)$ is a function of k and k_n is its set of roots. In the case of the δ-function above, the roots are the wave vectors k_f of the final state that conserve energy. For static perturbations, this is:

$$k_f = \sqrt{k_i^2 + \frac{2m^*}{\hbar^2}(E_i - E_f)} \tag{10.219}$$

while for sinusoidal perturbations, the boson energy must be included:

$$k_f = \sqrt{k_i^2 + \frac{2m^*}{\hbar^2}(E_i - E_f \pm \hbar\omega)} \tag{10.220}$$

where the upper sign represents absorption and the lower sign represents emission. Note that the negative solution for k_f has been ignored here, because k_f is the *magnitude* of the final wave vector, and can therefore only take a positive value.

The scattering rate can now be rewritten in terms of the energy-conserving final wave vector:

$$W_{if}(k_i, k_\alpha) = \frac{2\pi m^*}{\hbar^3 k_f} \left|\langle f|\tilde{\mathcal{H}}|i\rangle\right|^2 \delta(k_\alpha - k_f)\Theta\left(k_\alpha^2\right) \tag{10.221}$$

The rate at which a carrier in subband i with wave vector k_i scatters to *any* state within subband f is then given by summing over all states in the final subband:

$$W_{if}(k_i) = \frac{2\pi m^*}{\hbar^3 k_f} \sum_{k_\alpha} \left|\langle f|\tilde{\mathcal{H}}|i\rangle\right|^2 \delta(k_\alpha - k_f)\Theta\left(k_\alpha^2\right) \tag{10.222}$$

The sum over final states can be converted into an integral, multiplied by the area occupied by each state $A/(2\pi)^2$ (see Section 2.4):

$$W_{if}(k_i) = \frac{Am^*}{2\pi\hbar^3 k_f} \int_0^\infty \int_0^{2\pi} \left|\langle f|\tilde{\mathcal{H}}|i\rangle\right|^2 \delta(k_\alpha - k_f)\Theta\left(k_\alpha^2\right) k_\alpha \, \mathrm{d}k_\alpha \, \mathrm{d}\theta \tag{10.223}$$

where θ is the direction of the final wave vector in k space.

Now the integral over the final wave vector can be eliminated by using the δ-function to select only the energy-conserving value:

$$W_{if}(k_i) = \frac{Am^*}{2\pi\hbar^3}\Theta\left(k_f^2\right) \int_0^{2\pi} \left|\langle f|\tilde{\mathcal{H}}(k_f, \theta)|i\rangle\right|^2 \mathrm{d}\theta \tag{10.224}$$

Note that the wave-vector dependence of the scattering matrix element has been made explicit here. To summarise, the total scattering rate from a given initial state to all possible final

states within a subband is obtained by (i) finding the scattering matrix element at the energy-conserving final wave vector and (ii) integrating this over all angles. The lifetime of a given state $|i\rangle$ is then obtained by summing this over all final subbands:

$$\frac{1}{\tau_i(k_i)} = \sum_f W_{if}(k_i) \tag{10.225}$$

10.25 Impurity scattering

The interaction between carriers and impurities has been considered in two complementary ways so far. First, the *cumulative effect* of charged dopant ions (and the associated additional carriers) upon the band edge of a heterostructure can be determined by solving Poisson's equation, as described in Section 3.23. This approach treats the impurities as being part of a homogeneous volume of charge rather than considering them as individual particles, enabling large-scale 'band-bending' effects to be determined efficiently using the methods of classical physics. Second, the *microscopic* interaction between a single dopant ion and a single carrier within a heterostructure can be determined by solving the Schrödinger equation for such a system, as described in Chapter 5. This enables impurity states to be computed efficiently within lightly doped systems by assuming only a single impurity lies within the region of interest. In fact, the two models could be put together by first considering the 'background effect' of the numerous dopant ions upon the band edge, and subsequently solving the Schrödinger equation for a carrier interacting with a particular single impurity within the system.

However, the numerous dopant ions in the system *will* have some direct effect on a carrier in addition to the indirect band-bending. Classically, a carrier travelling through the system would encounter the Coulomb potential of the individual randomly distributed ions and this would cause some change in its momentum (i.e. scattering). On a quantum scale, it would be unfeasible to construct a Hamiltonian that models such a complex system. However, the scattering effect can be handled by solving Fermi's golden rule as before.

As was the case for carrier–carrier scattering, the perturbation to the total energy of the carrier is given by the Coulomb potential. It is assumed here that the carrier has opposite charge to the impurity (i.e. an electron interacting with a donor or a hole interacting with an acceptor), since this is the typical case in a doped semiconductor. The perturbation *for a single impurity* is then determined using:

$$\tilde{\mathcal{H}}_{\text{imp}}(\mathbf{r}-\mathbf{r_0}, z-z_0) = -\frac{e^2}{4\pi\epsilon\sqrt{|\mathbf{r}-\mathbf{r_0}|^2+(z-z_0)^2}}, \tag{10.226}$$

where $(z, \mathbf{r})$ and $(z_0, \mathbf{r_0})$ are the carrier and impurity positions, respectively. As before, note that the in-plane location has been separated from the location along the z-axis.

The scattering matrix element is obtained in a similar manner to the carrier–carrier interaction, except that it is only necessary to evaluate the integral with respect to carrier location since the impurity location is fixed:

$$\langle f|\tilde{\mathcal{H}}_{\text{imp}}|i\rangle = -\frac{e^2}{4\pi\epsilon A}\int \psi_f^*(z)\psi_i(z)\iint \frac{\mathrm{e}^{i(\mathbf{k}_f-\mathbf{k}_i)\bullet\mathbf{r}}}{\sqrt{|\mathbf{r}-\mathbf{r}_0|^2+(z-z_0)^2}}\,\mathrm{d}\mathbf{r}\,\mathrm{d}z \tag{10.227}$$

Making the substitutions $\mathbf{u} = \mathbf{r} - \mathbf{r_0}$ and $\mathbf{q} = \mathbf{k_i} - \mathbf{k_f}$, the scattering potential matrix element for a single impurity may be rewritten as

$$\langle f|\tilde{\mathcal{H}}_{\text{imp}}|i\rangle = -\frac{e^2}{4\pi\epsilon A}\mathrm{e}^{-i\mathbf{q}_\bullet\mathbf{r_0}} \int \psi_f^*(z)\psi_i(z) \iint \frac{\mathrm{e}^{-i\mathbf{q}_\bullet\mathbf{u}}}{\sqrt{|\mathbf{u}|^2 + (z-z_0)^2}}\ \mathrm{d}\mathbf{u}\,\mathrm{d}z \quad (10.228)$$

As was the case for carrier–carrier scattering, the two-dimensional integral over $\mathbf{u}$ is the Fourier transform of the Coulomb potential [33], and can be solved analytically to give:

$$\langle f|\tilde{\mathcal{H}}_{\text{imp}}|i\rangle = -\frac{e^2}{2\epsilon A q}\mathrm{e}^{-i\mathbf{q}_\bullet\mathbf{r_0}} \int \psi_f^*(z)\psi_i(z)\ \mathrm{e}^{-q|z-z_0|}\,\mathrm{d}z \quad (10.229)$$

Unlike carrier–carrier scattering, however, there is no need to account for momentum conservation within the system, since the impurity ions are many thousands of times more massive than the carrier. Now, the squared magnitude of this expression is:

$$\left|\langle f|\tilde{\mathcal{H}}_{\text{imp}}|i\rangle\right|^2 = \left(\frac{e^2}{2\epsilon A q}\right)^2 |I(q,z_0)|^2 \quad (10.230)$$

where

$$I(q,z_0) = \int \psi_f^*(z)\psi_i(z)\ \mathrm{e}^{-q|z-z_0|}\,\mathrm{d}z \quad (10.231)$$

Note that since the expression has been squared here, the sign in the Coulombic interaction is not important. In other words, the same scattering effect would result from either an attractive potential (e.g. an electron and a donor) or a repulsive potential (e.g. an electron and an acceptor).

The next task is to determine the cumulative effect of the numerous impurities in a system, which are distributed according to a spatially varying concentration, $d(z_0)$. The number of impurities in a thin slice of width $\mathrm{d}z_0$ is given by multiplying the concentration in the slice by the incremental width and the area of the system A. Now the total scattering matrix element is obtained by integrating equation (10.230) over the length of the system [34]:

$$\left|\langle f|\tilde{\mathcal{H}}|i\rangle\right|^2 = \frac{e^4}{4\epsilon^2 A q^2}\int d(z_0)\,|I(q,z_0)|^2\ \mathrm{d}z_0 \quad (10.232)$$

The scattering rate is now obtained by substituting this into equation (10.224):

$$W_{if}(k_i) = \frac{Am^*}{2\pi\hbar^3}\Theta\left(k_f^2\right)\int_0^{2\pi}\frac{e^4}{4\epsilon^2 A q^2}J_{if}(q)\,\mathrm{d}\theta \quad (10.233)$$

where the magnitude of the scattering wave vector is found using the cosine rule, $q^2 = k_f^2 + k_i^2 - 2k_i k_f\cos\theta$, and the following substitution has been made:

$$J_{if}(q) = \int d(z_0)\,|I(q,z_0)|^2\ \mathrm{d}z_0 \quad (10.234)$$

Finally, tidying up the constants in equation (10.233), and noting that the integrand is even-symmetric, yields:

$$W_{if}(k_i) = \frac{m^* e^4}{4\pi\hbar^3\epsilon^2}\Theta\left(k_f^2\right)\int_0^{\pi}\frac{J_{if}(q)}{q^2}\,\mathrm{d}\theta \quad (10.235)$$

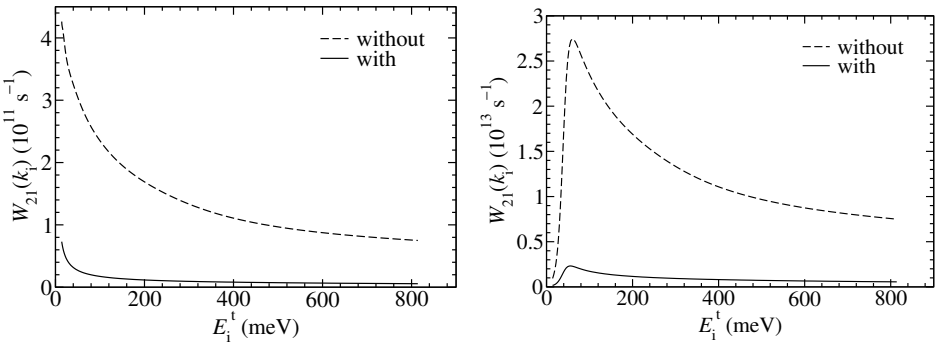


Figure 10.30: The impurity 2–1 scattering rate as a function of the initial electron energy, with and without screening, for a carrier density of $1 \times 10^{10}\,\mathrm{cm}^{-2}$ (left) and $100 \times 10^{10}\,\mathrm{cm}^{-2}$ (right) in each subband

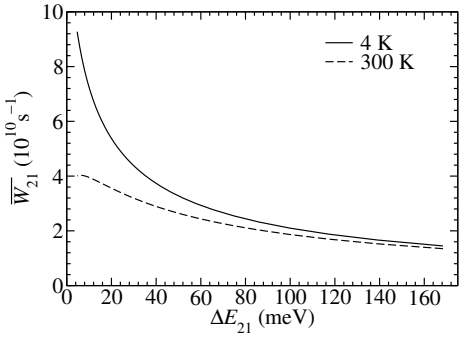


Figure 10.31: The mean 2–1 impurity scattering rate in an infinitely deep GaAs quantum well, as a function of the subband separation, including screening and final-state blocking. A carrier density of $1 \times 10^{10}\,\mathrm{cm}^{-2}$ was used in each subband. Band-bending effects have been omitted for simplicity

The intersubband scattering rate can then be averaged over the initial subband distribution, as described in Section 10.4.

This is a much simpler expression than equation (10.199) for carrier–carrier scattering, but similarities are clearly apparent. The prefactor is identical except for a scaling factor of 4π, and the integral expression is very similar in format (i.e. a form factor divided by the scattering wave vector). This reflects the fact that both scattering processes are derived from the Coulomb potential, with the only difference being whether the scatterer is a 'fixed' impurity or a second carrier. Indeed, although the impurity scattering is much faster to compute (since there is no need to integrate over a second carrier position), the same optimisation techniques can be applied as those described in Section 10.18. Namely, the form factor can be determined using a look-up table over q instead of recalculating it

for every combination of k_f and θ, and the I_{if} matrix element can be calculated using the computationally efficient form in equation (10.211).

As with carrier–carrier and LO phonon interactions, screening effects can be significant, since free carriers in the system are attracted to the impurities, and reduce the magnitude of the perturbation. Davies [1] shows that this effect may be included quite simply for impurity scattering by making the substitution $q \rightarrow q + q_{\mathrm{TF}}$ in the denominator of equation (10.235) where the Thomas–Fermi screening vector is given by:

$$q_{\mathrm{TF}} = \frac{m^* e^2}{2\pi\epsilon\hbar^2} \tag{10.236}$$

As was the case for carrier–carrier scattering, this substitution is advantageous because the singularity at $q = 0$ is removed.

Figure 10.30 shows the 2–1 scattering rate as a function of the initial electron energy, for an infinitely deep well with carrier densities of $1 \times 10^{10}\,\mathrm{cm}^{-2}$ (left) and $100 \times 10^{10}\,\mathrm{cm}^{-2}$ (right) in both subbands. In each case, impurities were distributed evenly throughout the well region with a sheet density equal to the total carrier density in both subbands (i.e. assuming complete ionisation of dopants). For simplicity, the band-bending effects that arise from the charge distribution were omitted. These plots show broadly the same characteristics as for the equivalent carrier–carrier interactions in Fig. 10.22, in that the rates tend to decrease as the initial carrier energy becomes very large, and increase approximately linearly with respect to the doping (or carrier) density.

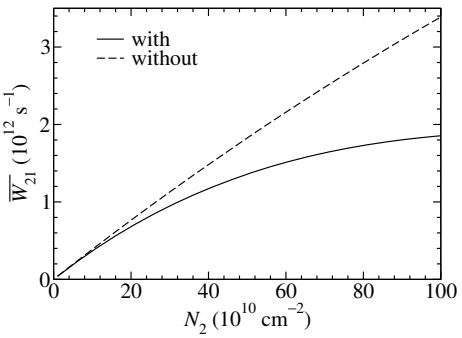


Figure 10.32: The mean 2–1 impurity scattering rate in a 400 Å infinitely deep GaAs quantum well, as a function of the carrier density in each subband. Results are shown both with and without the inclusion of final-state blocking effects. Band-bending effects have been omitted for simplicity.

However, there are several important distinctions. First, the effects of final-state blocking become apparent when the dopant density is very large, as in Fig. 10.30 (right), in which the scattering rate drops significantly for low initial subband energies. This behaviour is also apparent when the intersubband scattering rate is averaged over the initial carrier distribution, as shown in Fig. 10.31. Here, the final-state blocking is shown to become significant at high

temperatures, when the subband spacing is small. The effect of increasing the doping density is shown in Fig. 10.32. Without final-state blocking, the scattering rate increases almost linearly with respect to the carrier density. However, once the occupation of the final subband is taken into account, the linear relationship breaks down and the rate approaches a maximum value of $\sim 2 \times 10^{12}\,\mathrm{s}^{-1}$ at high carrier densities.

A second important feature is that the overall impurity scattering rates appear to be approximately an order of magnitude larger than the equivalent carrier–carrier scattering rate. However, this is not a straightforward comparison to make, because the impurity scattering rate also depends upon the *distribution* of impurities in the system. Indeed, as discussed in Section 3.26, there are several choices available for distributing dopants within a system.

Accordingly, Fig. 10.33 shows that the distribution of dopants within a system has a very strong effect on the scattering rate, whereas temperature plays a fairly weak role. The scattering rate is approximately ten times smaller when the dopant ions are placed within the barriers (i.e. modulation doping) rather than in the well. This is caused by the much smaller overlap of the impurity distribution with the electron states, leading to a smaller scattering matrix element in equation (10.234). This illustrates an advantage of modulation-doping designs, particularly in optical devices, in which very long carrier lifetimes are desirable in states that are involved in photon emission or absorption. In such devices, dopants are often positioned well away from the optically active part of the system, in order to minimise leakage currents. The figure shows a third possibility for dopant distribution, in which the carriers are all located within a very narrow slab within the centre of the well. In this 'δ-doping' configuration, the scattering rate is found to be around 40 times lower still, compared with the modulation-doped structure. This is because the wave function of the second state in a quantum well has a node (i.e. zero probability density) at the centre of the well. By positioning the dopants around this location, the overlap integral between the dopants and the 2–1 transition is heavily reduced. It is worth noting, however, that these symmetry 'tricks' only work in symmetric structures: if an electric field were applied to the well, for example, the node would shift away from the centre, and the scattering rate would increase. Nevertheless, the selection of dopant profile has been shown to be an important parameter in the scattering behaviour of systems.

10.26 Alloy disorder scattering

So far, the scattering events that have been considered have arisen from interactions between a carrier and another particle (i.e. a phonon, another carrier, or an impurity). However, it is important to recall that scattering arises from *any perturbation* to the system. One such perturbation is the random distribution of atoms within an alloy material (e.g. $\mathrm{Ga}_{1-x}\mathrm{Al}_x\mathrm{As}$). In pure semiconductor elements and compounds, the crystal potential is perfectly periodic, and the Bloch wave model gives an accurate representation of the system. This has underpinned the use of an envelope function approximation in this work, so far, in which the 'envelope' of a wave function was found by using a constant band-edge potential in the system Hamiltonian, for the entire layer of each material, rather than considering the microscopic periodic crystal potential.

However, in alloys (i.e. mixtures of two or more semiconductors), the atoms are distributed randomly, and as such the crystal potential is no longer strictly periodic. A 'virtual crystal'

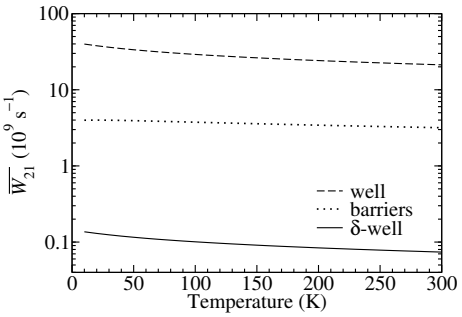


Figure 10.33: The mean 2–1 impurity scattering rate, as a function of temperature, for a range of dopant distributions within a $400\,$Å GaAs/$Ga_{0.85}Al_{0.15}As$ quantum well, with a sheet doping density of $1 \times 10^{10}\,\mathrm{cm}^{-2}$ in each case. Results are shown for doping spread evenly through the well, the barriers (i.e. modulation doping), or a narrow $10\,$Å slab in the centre of the well (i.e. δ-doping). Band-bending effects have been omitted for simplicity

model is generally used to approximate the properties of an alloy, by interpolating between the values for the pure materials (see Appendix A). Although this gives excellent results for the *static* behaviour of the system, i.e. the envelopes of wave functions and the energies of states can be computed very accurately, it ignores the random fluctuations in the crystal potential that occur on the atomic scale.

In classical terms, this random disorder of the crystal potential causes microscopic attractive or repulsive regions within the crystal, which give rise to scattering of carriers. The scattering rate can be evaluated using Fermi's golden rule, by determining the average (mean-square) magnitude of this random perturbation. First, the true (random) crystal potential could be determined by divided the system into pure-material primitive units at random locations. In zinc blende structures, e.g. GaAlAs alloys, these would be diatomic units (GaAs or AlAs), while in elemental, diamond-lattice structures, e.g. SiGe alloys, these would be single atoms (Si or Ge). Taking the example of a GaAlAs alloy, the total potential at position $\mathbf{R}$ would be given by:

$$V(\mathbf{R}) = \sum_{\mathbf{R}_i \in R_{\mathrm{GaAs}}} V_{\mathrm{GaAs}}(\mathbf{R} - \mathbf{R}_i) + \sum_{\mathbf{R}_i \in R_{\mathrm{AlAs}}} V_{\mathrm{AlAs}}(\mathbf{R} - \mathbf{R}_i) \tag{10.237}$$

where $V_{\mathrm{GaAs}}(\mathbf{R})$ is the potential around a GaAs diatomic unit, and $V_{\mathrm{AlAs}}(\mathbf{R})$ is the potential around an AlAs unit. These are summed over the sets of all the random locations R_{GaAs} and R_{AlAs} that contain the respective diatomic units.

This expression, therefore, takes account of the random distribution of atoms in the alloy. In contrast, the virtual crystal approach approximates every unit in the system with an identical, average potential that is interpolated between the values of the pure materials:

$$V_{\mathrm{vc}}(\mathbf{R}) = xV_{\mathrm{AlAs}}(\mathbf{R}) + (1-x)V_{\mathrm{GaAs}}(\mathbf{R}) \tag{10.238}$$

where x is the mole fraction of Al in the $Ga_{1-x}Al_xAs$ alloy. In this model, the total potential at a given point is then given by summing the alloy potential over *all* the diatomic unit

locations, R_{A}:

$$V'(\mathbf{R}) = \sum_{\mathbf{R}_i \in R_{\mathrm{A}}} V_{\text{alloy}}(\mathbf{R} - \mathbf{R}_i) \tag{10.239}$$
$$= \sum_{\mathbf{R}_i \in R_{\mathrm{A}}} [x V_{\text{AlAs}}(\mathbf{R} - \mathbf{R}_i) + (1-x) V_{\text{GaAs}}(\mathbf{R} - \mathbf{R}_i)]$$

Note that since the unit locations R_{A} are periodic, and the potential around each location is identical, the virtual crystal potential is also a purely periodic function.

The random perturbation to carriers in the system is then given by the difference between the true (random) crystal potential and the virtual crystal approximation:

$$\tilde{\mathcal{H}}(\mathbf{R}) = \sum_{\mathbf{R}_i \in R_{\text{AlAs}}} (1-x)\Delta V(\mathbf{R} - \mathbf{R}_i) - \sum_{\mathbf{R}_i \in R_{\text{GaAs}}} x \Delta V(\mathbf{R} - \mathbf{R}_i) \tag{10.240}$$

where $\Delta V(\mathbf{R}) = V_{\text{AlAs}}(\mathbf{R}) - V_{\text{GaAs}}(\mathbf{R})$ is the difference between the GaAs and AlAs diatomic potentials.

Now, a simplification can be made by assuming that the diatomic potentials are extremely narrow (non-overlapping) functions. The difference in potentials is then given by:

$$\Delta V(\mathbf{R}) \approx \Omega \Delta V_{\text{ad}} \delta(\mathbf{R}) \tag{10.241}$$

where Ω is the volume containing a single pure-material unit and ΔV_{ad} is the spatial average of the difference in potential between the two materials over that volume [35]. There is, therefore, some freedom in the choice of volume, Ω in this expression, since the value of the disorder potential ΔV_{ad} is obtained by averaging over whichever volume is used. In other words, a larger volume would result in a smaller disorder potential, and vice versa [36]. For a zinc blende structure, each unit cell of the crystal contains four diatomic primitive cells, and hence $\Omega = A_0^3/4$ is a common choice, where A_0 is the lattice constant of the material [37]. Conversely, in an elemental (diamond-like) system, each monatomic primitive unit occupies a volume $\Omega = A_0^3/8$ and this is, therefore, often selected as the basis for SiGe alloys [36]. The value of the disorder potential could be obtained from first principles using electronic-structure calculations such as those described in Chapter 15 to determine the average offset between the two crystal potentials within a small volume [38]. Alternatively, the bulk band offset between the two pure materials could be used as a first approximation. However, in practice, the value is obtained experimentally for a given choice of Ω. In the following calculations, the values $\Omega = A_0^3/4$ and $\Delta V_{\text{ad}} = 0.6\,\text{eV}$ are taken from experimental values for GaAlAs alloys, as previously used by Yoon *et al*. [37].

A final expression for the random perturbation is now found by substituting equation (10.241) into equation (10.240) to obtain:

$$\tilde{\mathcal{H}}(\mathbf{R}) = \Omega \Delta V_{\text{ad}} \left[\sum_{\mathbf{R}_i \in R_{\text{AlAs}}} (1-x)\delta(\mathbf{R} - \mathbf{R}_i) - \sum_{\mathbf{R}_i \in R_{\text{GaAs}}} x \delta(\mathbf{R} - \mathbf{R}_i) \right] \tag{10.242}$$

The scattering matrix element is then found, as usual, by deriving the overlap of this perturbation with the initial and final wave functions:

$$\langle f|\tilde{\mathcal{H}}|i\rangle = \frac{\Omega\Delta V_{\text{ad}}}{A}\iiint \mathrm{d}\mathbf{R}\ \psi_f^*(z)\psi_i(z)\mathrm{e}^{-\mathrm{i}\mathbf{q}\cdot\mathbf{r}} \tag{10.243}$$

$$\times\left\{\sum_{\mathbf{R}_i\in R_{\text{AlAs}}}(1-x)\delta(\mathbf{R}-\mathbf{R}_i) - \sum_{\mathbf{R}_i\in R_{\text{GaAs}}} x\delta(\mathbf{R}-\mathbf{R}_i)\right\}$$

where the substitution $\mathbf{q} = \mathbf{k}_i - \mathbf{k}_f$ has been used (as in Section 10.25).

Now the Dirac δ-functions can be used to eliminate the integrals, since they only have a non-zero value at each of the locations of the diatomic units:

$$\langle f|\tilde{\mathcal{H}}|i\rangle = \frac{\Omega\Delta V_{\text{ad}}}{A}\sum_{\mathbf{R}_i\in R_{\text{AlAs}}}[1-x(z_i)]\,\psi_f^*(z_i)\psi_i(z_i)\mathrm{e}^{-\mathrm{i}\mathbf{q}\cdot\mathbf{R}_i} \tag{10.244}$$

$$-\frac{\Omega\Delta V_{\text{ad}}}{A}\sum_{\mathbf{R}_i\in R_{\text{GaAs}}} x(z_i)\psi_f^*(z_i)\psi_i(z_i)\mathrm{e}^{-\mathrm{i}\mathbf{q}\cdot\mathbf{R}_i} \tag{10.245}$$

Note that the spatial dependence of the alloy composition has now been made explicit, since the focus of this work is on two-dimensional systems.

The complete scattering matrix element is therefore found by summing over all locations $\mathbf{R}_i$ and selecting either the first or second term in equation (10.245) depending on whether an AlAs or a GaAs unit lies at that location. However, this clearly requires knowledge of the exact location of every randomly distributed unit in the system. Instead, the *mean-square* value can be obtained by averaging over all locations [35]:

$$|\langle f|\tilde{\mathcal{H}}|i\rangle|^2 = \frac{\Omega\Delta V_{\text{ad}}^2}{A}\int \mathrm{d}z|\psi_f(z)|^2x(z)[1-x(z)]|\psi_i(z)|^2 \tag{10.246}$$

The scattering rate within a two-dimensional system can be found by substituting this into equation (10.224):

$$W_{if}(k_i) = \frac{Am^*}{2\pi\hbar^3}\Theta\left(k_f^2\right)\int_0^{2\pi}\frac{\Omega\Delta V_{\text{ad}}^2}{A}\int|\psi_f(z)|^2x(z)[1-x(z)]|\psi_i(z)|^2\mathrm{d}\theta\mathrm{d}z \tag{10.247}$$

Noting that the scattering matrix element is independent of angle, the integral over θ can be separated and evaluates to a constant 2π. Finally, simplifying the coefficients yields:

$$W_{if}(k_i) = \frac{m^*\Omega\Delta V_{\text{ad}}^2}{\hbar^3}\Theta\left(k_f^2\right)\int|\psi_f(z)|^2x(z)[1-x(z)]|\psi_i(z)|^2\mathrm{d}z \tag{10.248}$$

As usual, the intersubband scattering lifetimes can be computed by averaging this expression over all initial states, as described in Section 10.4. It is worth noting that this expression is independent of the initial wave vector, provided that the resulting k_f is real. As such, the alloy disorder scattering rate is constant with respect to the initial carrier energy. In other words, *all* carriers in the initial subband are scattered equally rapidly by the alloy disorder and the rate depends only upon the overlap between the probability density of carriers in each subband and the composition of the alloy material. The lack of wave-vector dependence also means that the scattering rate can be computed very rapidly, with no need for any computational optimisation techniques on modern computer hardware.

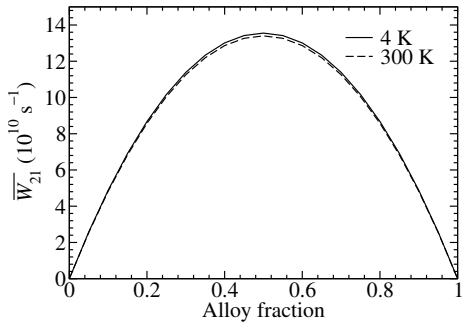


Figure 10.34: The mean 2–1 intersubband alloy-disorder scattering rate for a $400\,\text{Å}$ infinitely deep $Ga_{1-x}Al_xAs$ quantum well with varying alloy composition. An electron density of $1 \times 10^{10}\,\text{cm}^{-2}$ was used in each subband. For simplicity, a constant effective mass of $m^* = 0.067m_0$ has been used for all materials

10.27 Alloy disorder scattering in quantum wells

The alloy disorder scattering rate in a quantum well structure is highly dependent on the term $x(1-x)$ in equation (10.248). This term sets the rate to zero in pure materials (i.e. $x = 0$ or $x = 1$), and as such there is no alloy scattering within a heterostructure consisting of two pure semiconductors (e.g. a GaAs well with AlAs barriers). Figure 10.34 shows the scattering rate within an infinitely deep $Ga_{1-x}Al_xAs$ well with varying alloy composition. The rate reaches a maximum value of $1.4 \times 10^{11}\,\text{s}^{-1}$ when the alloy is evenly mixed (i.e. $x = 0.5$), and is shown to be almost temperature-independent. Even at relatively modest alloy compositions, the scattering rate is comparable to the impurity, carrier–carrier and deformation-potential scattering rates considered earlier in the chapter. For example, a $Ga_{0.9}Al_{0.1}As$ well material gives rise to a scattering rate of $5 \times 10^{10}\,\text{s}^{-1}$. As such, alloy disorder scattering can be a dominant process in alloy materials, particularly at low temperatures, when other processes are reduced. Indeed, this provides some motivation for constructing quantum wells using a pure semiconductor in the well region and an alloy within the barriers, rather than the other way around.

In the latter case, the scattering rate is much lower, since only the 'tails' of the wave functions penetrate into the alloy barriers, while most of the electron probability remains in the pure material within the well. The scattering rate for such a system is plotted in Fig. 10.35, and is shown to be significantly lower. Interestingly, the scattering rate is no longer maximised at $x = 0.5$, because higher alloy fractions give rise to higher barrier potentials, and the wave functions become much more strongly confined within the well. As such, the overlap between the wave function 'tails' and the alloy reduces very strongly as the alloy fraction increases. The conclusion is, therefore, that alloy disorder scattering is minimised in systems with a pure material in the well and a large alloy fraction in the barriers. Figure 10.35 also shows that the intrasubband scattering rate is approximately double that of the intersubband rate, in line with the results found for other processes.

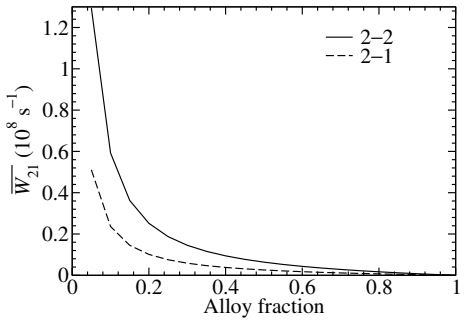


Figure 10.35: The mean 2–2 and 2–1 alloy disorder scattering rates for a $400\,$Å GaAs quantum well, surrounded by a pair of $200\,$Å $Ga_{1-x}Al_xAs$ barriers with varying alloy composition. An electron density of $1 \times 10^{10}\,\mathrm{cm}^{-2}$ was used in each subband, with a fixed temperature of $4\,$K. For simplicity, a constant effective mass of $m^* = 0.067m_0$ has been used for all materials

10.28 Interface roughness scattering

So far, the assumption has been made that two-dimensional systems have a potential profile that varies only in the z-direction. In other words, the grown heterostructures are assumed to consist of a stack of perfectly smooth and flat layers. This assumption allows the wave function to be split into two parts: completely free-carrier motion along the plane, and a confined mode in the z-direction, which can be found by solving the Schrödinger equation for the one-dimensional potential $V(z)$. In reality, though, the growth of heterostructures is not perfect. For example, Chapter 4 considered the diffusion of alloy components across an interface in the z-direction.

Another very significant effect is the roughness along the plane, i.e. random variations in the thickness of each layer over the x–y plane. In classical terms, the wider or narrower regions within each layer will respectively enhance or restrict carrier motion (i.e. attractive or repulsive potentials), giving rise to a scattering effect. As usual, the lifetimes within the quantum-confined system may be determined by solving Fermi's golden rule. The fluctuations in interface location are usually assumed to have a Gaussian Fourier transform $\Delta_z(\mathbf{r})$ such that

$$\langle \Delta_z(\mathbf{r}) \Delta_z(\mathbf{r}') \rangle = \Delta^2 \exp\left(-\frac{|\mathbf{r} - \mathbf{r}'|}{\Lambda^2} \right) \tag{10.249}$$

where Δ is the mean height of the roughness profile and Λ is a correlation length, i.e. a measure of the length scale in x, y over which the interface varies [28, 39, 40]. Values of 3–5 Å and 50–70 Å have been obtained experimentally for Δ and Λ in GaAs-based heterostructures, respectively [39].

Most models of interface roughness scattering consider perfectly abrupt interfaces between two materials (i.e. a step-like potential), [41], and good agreement with experimental lifetimes for such structures has been obtained [42]. However, a more general model [43] can be applied to arbitrary interface geometries, such as the diffuse profiles considered

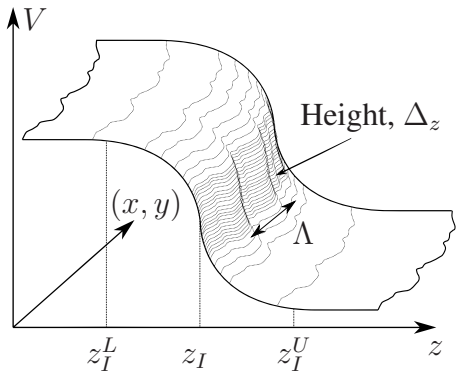


Figure 10.36: Schematic illustration of the potential profile around a diffuse interface, with a roughness profile similar to that in equation (10.249)

in Chapter 4. Following this approach, Fig. 10.36 illustrates a potential profile, consisting of an interface at position z_I, after alloy components have been allowed to diffuse across the interface. The upper and lower extent of the diffuse profile are denoted z_I^{U} and z_I^{L}, respectively. In principle, these should be set infinitely far from the interface location, but in heterostructures with multiple interfaces they can be set to either a fixed distance (e.g. a few multiples of the diffusion length) or halfway between adjacent interfaces.

The figure shows the effect of in-plane roughness, leading to a set of random bumps of height $\Delta_z(\mathbf{r})$. In effect, this shifts the potential profile at each in-plane location from $V(z)$ to $V[z - \Delta_z(\mathbf{r})]$. The perturbation in the potential around the Ith interface is then given by subtracting one from the other:

$$\tilde{\mathcal{H}}_I(\mathbf{R}) = \mathrm{rect}\left(\frac{z - z_I}{z_I^{\mathrm{U}} - z_I^{\mathrm{L}}}\right)\{V[z - \Delta_z(\mathbf{r})] - V(z)\} \tag{10.250}$$

The rectangular ('rect') function here simply sets the perturbing potential to zero at all positions $z \geq z_I^{\mathrm{U}}$ and $z < z_I^{\mathrm{L}}$ so that each interface is considered independently.

The expression can be simplified by noting that the perturbing potential has a very narrow width Δ_z, and as such, it can be approximated using:

$$\tilde{\mathcal{H}}_I(\mathbf{R}) = -\Delta_z(\mathbf{r})\frac{\mathrm{d}V(z)}{\mathrm{d}z}\,\mathrm{rect}\left(\frac{z - z_I}{z_I^{\mathrm{U}} - z_I^{\mathrm{L}}}\right) \tag{10.251}$$

The contribution of the Ith interface to the scattering matrix element is then given, as usual, by integrating the overlap with the carrier wave functions:

$$\langle f|\tilde{\mathcal{H}}_I|i\rangle = -\frac{F_{if,I}}{A}\iint \mathrm{e}^{-i\mathbf{q}\cdot\mathbf{r}}\Delta_z(\mathbf{r})\,\mathrm{d}\mathbf{r} \tag{10.252}$$

where $\mathbf{q} = \mathbf{k}_i - \mathbf{k}_f$ as before, and

$$F_{if,I} = \int_{z_I^{\mathrm{L}}}^{z_I^{\mathrm{U}}} \psi_f^*(z)\frac{\mathrm{d}V(z)}{\mathrm{d}z}\psi_i(z)\,\mathrm{d}z \tag{10.253}$$

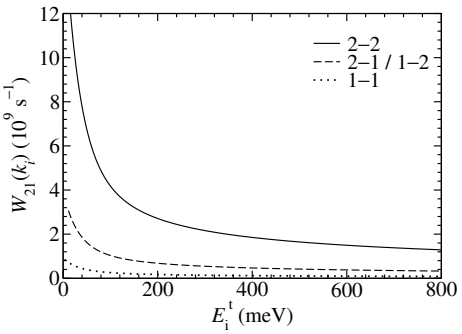


Figure 10.37: The 2–2, 2–1, 1–2 and 1–1 interface roughness scattering rates as a function of initial electron energy in a $400\,\text{Å}$ GaAs/$Ga_{0.85}Al_{0.15}As$ quantum well. A carrier density of $1 \times 10^{10}\,\text{cm}^{-2}$ in each subband was used, and the rates are independent of temperature. The roughness parameters were set as $\Delta = 3\,\text{Å}$ and $\Lambda = 50\,\text{Å}$

As was the case with alloy disorder scattering, the random distribution of the roughness is accounted for by taking the mean-square of the scattering matrix element:

$$|\langle f|\tilde{\mathcal{H}}_I|i\rangle|^2 = \left(\frac{F_{if,I}}{A}\right)^2 \iint \mathrm{d}^2\mathbf{r} \iint \mathrm{d}^2\mathbf{r}'\Delta_z(\mathbf{r})\Delta_z(\mathbf{r}')\mathrm{e}^{-i\mathbf{q}\cdot\mathbf{u}} \tag{10.254}$$

$$= \frac{\left(F_{if,I}\Delta\Lambda\right)^2\pi}{A}\mathrm{e}^{-q^2\Lambda^2/4}$$

where $\mathbf{u} = \mathbf{r} - \mathbf{r}'$.

Substituting this into equation (10.224) gives the scattering rate within a two-dimensional system due to the perturbation at a single interface:

$$W_{if,I}(k_i) = \frac{\left(F_{if,I}\Delta\Lambda\right)^2 m^*}{\hbar^3}\Theta(k_f^2)\int_0^\pi \mathrm{e}^{-q^2\Lambda^2/4}\,\mathrm{d}\theta \tag{10.255}$$

The integral may be solved quasi-analytically by substituting $q^2 = k_i^2 + k_f^2 - 2k_ik_f\cos\theta$ and using equation (3.339) of Gradshteyn and Ryzhik [13]:

$$\int_0^\pi \mathrm{e}^{a\cos\theta}\,\mathrm{d}\theta = \pi I_0(a) \tag{10.256}$$

where $I_0(a)$ is the regular modified cylindrical Bessel function of zeroth order.

The scattering rate at a given interface is now

$$W_{if,I}(k_i) = \frac{\pi m^*(\Delta\Lambda)^2}{\hbar^3}\beta(k_i)\left(F_{if,I}\right)^2, \tag{10.257}$$

where

$$\beta(k_i) = \mathrm{e}^{-(k_i^2+k_f^2)\Lambda^2/4} I_0\left(\frac{k_ik_f\Lambda^2}{2}\right)\Theta(k_f^2) \tag{10.258}$$

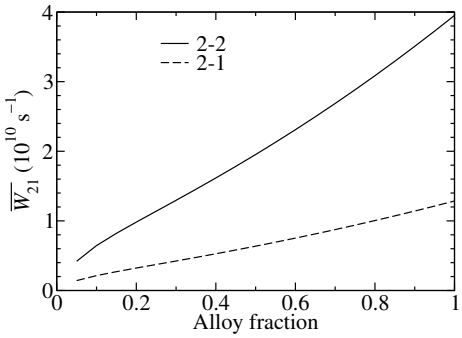


Figure 10.38: Mean 2–2 and 2–1 interface roughness scattering rates in a $400\,\text{Å}$ GaAs/$Ga_{1-x}Al_xAs$ quantum well as a function of alloy fraction. A carrier density of $1 \times 10^{10}\,\text{cm}^{-2}$ was used in each subband, and roughness parameters were set as $\Delta = 3\,\text{Å}$ and $\Lambda = 50\,\text{Å}$

At this stage, it is worth checking that this generalised model reduces to the step-like interface models that have typically been used in the literature. In these cases, the potential profile at the interface is a Heaviside step function $V(z) = V_0\Theta(z - z_I)$, and its derivative is the Dirac δ-function. Substituting this into equation (10.253) gives:

$$F_{if,I} = \int_{z_I^{\text{L}}}^{z_I^{\text{U}}} \psi_f^*(z) V_0 \delta(z - z_I) \psi_i(z)\,\mathrm{d}z \tag{10.259}$$

The δ-function can be used to eliminate the integral, finally yielding:

$$F_{if,I} = \psi_f^*(z_I) V_0 \psi_i(z_I) \tag{10.260}$$

which is the same as the expected, commonly used form [28].

Returning now to the general expression (equation (10.257)), all that remains is to find the total scattering rate for all interfaces in the system. The simplest approach is to assume that the roughness in each layer is uncorrelated, in which case equation (10.257) is simply summed over all interfaces:

$$W_{if}(k_i) = \frac{\pi m^*(\Delta\Lambda)^2}{\hbar^3}\beta(k_i)\sum_I \left(F_{if,I}\right)^2 \tag{10.261}$$

This is a reasonable approximation for lattice-matched systems such as GaAs/$Ga_{1-x}Al_xAs$, since good growth quality is achievable and strain effects are very weak. Other models have been proposed [40], which allow for correlation between the roughness in different layers, as may be the case in strained materials such as Si/SiGe.

Equation (10.261) can be solved rapidly without any requirement for computational optimisation techniques, since the $\beta(k_i)$ term contains the only wave-vector dependence, and has an analytical solution. It is also worth noting that $\beta(k_i)$ does not depend directly on the structure, but only on the separation between subbands. The $F_{if,I}$ terms contain all the

information about the potential profile and wave functions, but are wave-vector-independent and therefore only need to be computed once for each intersubband process. As usual, the average intersubband scattering rates can be computed by averaging equation (10.261) over the carrier distributions in each subband as described in Section 10.4. None of the terms in equation (10.261) depend on temperature, although a weak temperature dependence is introduced in the *average rates* through the effect on carrier distribution.

10.29 Interface roughness scattering in quantum wells

Figure 10.37 shows the scattering rate as a function of initial electron energy within a quantum well. The rate is shown to be largest for electrons near the band edge, since $\beta(k_i)$ is a monotonically decaying function of wave vector. Since the scattering rate depends upon the overlap of carrier wave functions at the interfaces, it is largest for transitions involving weakly bound excited states. As such, the 2–2 scattering rate is larger than the 2–1 (or 1–2) rate, which in turn is larger than the 1–1 rate.

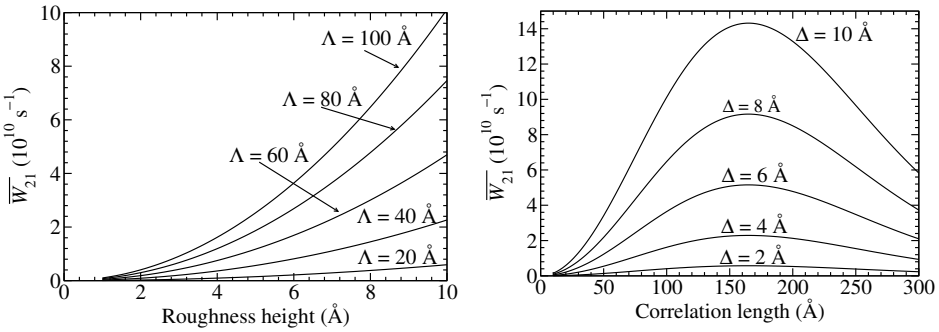


Figure 10.39: The mean 2–1 interface roughness scattering rate in a $400\,\text{Å}$ GaAs/$Ga_{0.85}Al_{0.15}As$ quantum well as a function of the roughness height Δ (left) and correlation length Λ (right). Results are shown for a fixed carrier density of $1 \times 10^{10}\,\text{cm}^{-2}$ in each subband and a temperature of $4\,\text{K}$

Figure 10.38 shows the effect of changing the barrier composition in a quantum well. The scattering rate is shown to increase monotonically as the $Ga_{1-x}Al_xAs$ alloy fraction increases, owing to the increasing potential step. This is in contrast with the results for alloy disorder scattering, in which the rate was found to decrease as the barriers grow. The reason for this is that interface roughness scattering depends only on the wave-function overlap very near to the interface, while alloy disorder scattering depends on the squared magnitude of the decaying tails of the wave function within the barriers. Furthermore, the interface roughness scattering rate is shown to be of the order of 100 times larger than the equivalent alloy disorder scattering rate in Fig. 10.35. As such, a system with somewhat lower barriers ($x \sim 15\%$) will yield the lowest total scattering rate.

The effect of the roughness distribution is shown in Fig. 10.39. The roughness height appears as a pre-factor Δ^2 in equation (10.261), and as such the total scattering rate increases

with the square of Δ in Fig. 10.39 (left). The $\beta(k_i)$ term within the scattering matrix element is a rather more complicated function of the correlation length, and Fig. 10.39 (right) shows that the scattering rate increases to a maximum around $\Lambda = 165\,\text{Å}$ before decreasing again at longer correlation lengths. Although the Bessel function in equation (10.258) has no general closed-form analytical solution, the origin of the peak in Fig. 10.39 (right) can be understood by making a simple approximation that at low temperatures all carriers lie very near the bottom of the subband, and as such $k_i \approx 0$. At this point, the argument of the Bessel function becomes zero, and the standard result $I_0(0) = 1$ can be used and the function simplifies to:

$$\beta(k_i) = \exp\left(-\frac{k_f^2\Lambda^2}{4}\right) \tag{10.262}$$

The scattering rate in equation (10.261) is then proportional to:

$$Y_{if} = \Lambda^2 \exp\left(-\frac{k_f^2\Lambda^2}{4}\right) \tag{10.263}$$

The maximum value can then be found by differentiating this with respect to Λ:

$$\frac{\mathrm{d}Y_{if}}{\mathrm{d}\Lambda} = \Lambda \exp\left(-\frac{k_f^2\Lambda^2}{4}\right)\left(2 - \frac{k_f^2\Lambda^2}{2}\right) \tag{10.264}$$

Setting this to zero, and noting that the final term is the only one with a solution between 0 and ∞, gives the correlation length at which the peak scattering rate occurs:

$$\Lambda_0 = \frac{2}{k_f} = \frac{2\hbar}{\sqrt{2m^*(E_i - E_f)}} \tag{10.265}$$

Substituting in the subband separation $E_i - E_f = 8.41\,\text{meV}$ for this system yields $\Lambda_0 = 164\,\text{Å}$ in close agreement with the numerical simulations in Fig. 10.39 (right).

In practice, the roughness correlation length is a fixed parameter, which is determined by the quality of the epitaxial growth of the system. It follows, though, that the $\beta(k_i)$ parameter will be enhanced when the subband separation matches the peak condition in equation (10.265), i.e. when $E_i - E_f = 4\hbar^2/(2m^*\Lambda^2)$. For $\Lambda = 50\,\text{Å}$, $\beta(k_i)$ is maximised when the subband separation is $90\,\text{meV}$, while for $\Lambda = 150\,\text{Å}$ a subband separation of $10\,\text{meV}$ is found. One might conclude, at first glance, that peaks in the total scattering rate could be observed at these subband separations by investigating finite quantum wells of different widths. However, this is not the complete picture, since the penetration of wave functions into the barriers also increases very significantly in finite quantum wells as the states are pushed to higher energies, and hence the $F_{if,I}$ term grows in equation (10.261). As such, the total scattering rate reaches its peak value at somewhat larger subband separations than the peak value of $\beta(k_i)$. Figure 10.40 shows, for example, that at a fixed subband spacing of $10\,\text{meV}$, $\Lambda_0 = 150\,\text{meV}$ as calculated above, since $\beta(k_i)$ takes its maximum value at this subband spacing. However, if the well is made narrower (hence increasing the subband separation), then the excited wave functions penetrate deeper into the barriers, and the scattering rate continues to grow. The peak in the scattering rate is then observed at a subband separation of $31\,\text{meV}$.

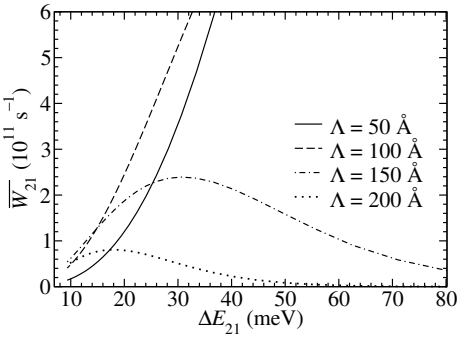


Figure 10.40: Mean 2–1 interface roughness scattering rate as a function of subband separation in a GaAs/AlAs quantum well with varying width. Results are shown for a fixed carrier density of $1 \times 10^{10}\,\mathrm{cm}^{-2}$ in each subband and a temperature of $4\,\mathrm{K}$

Clearly, this behaviour is a rather complex function of the structure being considered, and depends on the solutions of the Schrödinger equation, i.e. the relationship between the physical dimensions of the system, the subband separation and the decay constant for the wave functions within the barriers. A few general qualitative conclusions may be deduced, however. First, that the interface roughness scattering may be enhanced (or reduced) by adjusting the dimensions of the system. Second, that larger roughness correlation lengths generally lead to peaks in the total scattering rate at lower subband separations. Finally, that the subband separation that maximises $\beta(k_i)$ can be estimated using a simple analytical expression (equation (10.265)) and the total scattering rate is maximised at a subband separation a few times larger than this. Taking the relatively short correlation length ($\Lambda = 50\,\text{Å}$) reported for GaAs/$Ga_{1-x}Al_xAs$ systems, one can conclude that the peak scattering rate would not be seen until subband separations several times greater than $90\,\mathrm{meV}$ were reached in a heterostructure. This is rather large compared with the typical energy scales under consideration, and as such the simpler approximation can be made that interface roughness scattering increases monotonically with respect to subband separation in such systems.

As a final note, interface roughness scattering is clearly dependent on the number of interfaces within a system. In devices such as quantum cascade lasers, in which the wave functions overlap a large number of thin barriers, the interface roughness scattering rate can become quite large and must be accounted for carefully in the device design process.

10.30 Carrier scattering in quantum wires and dots

Carrier scattering in lower-dimensional systems, wires and dots, will be an important issue as they are introduced into future generations of optoelectronic devices. For the theoretical understanding of the carrier dynamics, scattering rate expressions will be needed.

Although the scattering rates derived here are not applicable to such systems, the techniques employed to obtain them do have some relevance. Quantum wires resemble quantum wells more closely than dots, as they still retain carrier dispersion. Therefore, developing a model of carrier scattering with *bulk* LO phonons will require expressing

the carrier wave function as a product of a two-dimensional envelope function with a one-dimensional plane wave, as opposed to the other way around. The remainder of the derivation should follow in a similar fashion. Similar arguments for the route to carrier–carrier scattering in quantum wires may also be applicable.

However, for quantum dots the situation is quite different as the carriers never possess dispersion; hence, for scattering with phonons of fixed energies, it would appear *a priori* that this will only occur for resonance conditions, i.e. when the sublevel separation is equal to a phonon energy. Carrier–carrier scattering between the sublevels of quantum dots resembles the Coulombic interaction in multi-electron atoms, with the latter being an area where much work has been done.

Theoretical, and hence computational studies of carrier scattering in quantum wires and dots are still quite rare in the literature (although see, for example, [44, 45, 46, 47, 48, 49]). For an introduction to some aspects of the particulars of optical processes (carrier–photon scattering) in wires and dots, see Basu ([50], p. 343).

Exercises

(1). (a) Plot the expected number of phonons per state (the Bose–Einstein factor N_0) as a function of temperature T for phonon energies of 1, 2, 3 and 4 meV.

(b) Plot the Bose–Einstein factor N_0 as a function of phonon energy $\hbar\omega$ for temperatures of 4, 10, 20, 77 and 300 K.

(c) By assuming a phonon displaces just a single pair of Ga and As atoms, derive an expression for the change in electrostatic potential experienced by an electron at a distance r from the bond centre of this isolated electric dipole. By choosing typical values for the material constants of GaAs, plot the change in electrostatic potential along the axis of the dipole and at angles of 30, 60 and 90° to it. (Take the ionicity fraction f_i of the GaAs bond as 0.4, i.e. 0.4 of an electron transfers from As to Ga when they bond; see [51], p. R153, and end-of-paper references.)

(d) Use the QWWAD codes to explore what happens to the thermally averaged intersubband phonon scattering rate as a function of the carrier density in the (lower-energy) final state of a quantum well. For simplicity base the calculations in a GaAs quantum well surrounded by infinitely high barriers.

In the first instance choose a quantum well width to give a separation between the ground and first excited states of about $6\,\mathrm{meV}$ and show the effect of 'final-state blocking' for electron–acoustic phonon scattering.

Repeat the calculations but with narrower wells to give separations between the ground and first-excited states of 30 and $40\,\mathrm{meV}$ and investigate the effect of the carrier density in the lower-energy final state on electron–LO phonon scattering.

Comment on the results.

(2). In the treatise here, the possibility of final-state blocking in electron–electron scattering is ignored as this really needs the final state momenta of the electrons to be calculated (see [19]). However, thermodynamics would suggest that if two electrons collide then generally the higher-energy electron will transfer kinetic energy to the lower-energy electron. So the approximation could be made that the electrons have the

same final-state energy. Making this simplification, show how the thermal averaging of intersubband electron–electron scattering (see Section 10.17) could be adapted to incorporate this. Illustrate this with the example of increasing electron density in the ground state of an infinitely deep GaAs quantum well, on the lifetime of the 22–11 electron–electron scattering mechanism.

(3). In Section 10.25 it was shown that impurity scattering between a pair of states can be minimised by placing donors in thin sheets at the nodes of the wave functions. Use this principle to design a $90\,\text{Å}$ GaAs/$Ga_{0.6}Al_{0.4}As$ quantum well with three bound states and a total sheet doping density of $1 \times 10^{10}\,\text{cm}^{-2}$, in which carriers in state $|3\rangle$ have a longer lifetime than those in state $|2\rangle$. Assume a carrier temperature of 77 K. Use the QWWAD simulation tools to confirm that $\tau_3 > \tau_2$.

(4). Find a solution for the alloy disorder scattering rate in an infinitely deep quantum well by solving equation (10.248) analytically. Assume a constant alloy fraction across the well, and let the initial subband index be greater than or equal to the final subband index, such that $\Theta(k_f) = 1$. Determine the ratio between intersubband and intrasubband scattering rates.

(5). Consider a double quantum well structure consisting of a pair of GaAs wells separated by a $Ga_{0.8}Al_{0.2}As$ barrier, which is subjected to an annealing process, such that the alloy composition diffuses between layers as described in Chapter 4. Take the well widths to be 60 Å and 40 Å, the barrier width to be 30 Å, and embed the system between two 200 Å barriers. Use a constant doping density of $1 \times 10^{15}\,\text{cm}^{-3}$ throughout the well regions and a carrier temperature of 77 K. Use the `qwwad_diffuse` program to simulate the diffuse structure for a range of annealing times, and compute its band structure using the appropriate QWWAD simulation tools. Assume a constant diffusion coefficient $D_0 = 10\,\text{Å}^2\,\text{s}^{-1}$. Calculate the mean alloy disorder and interface roughness scattering rates from the second to the first subband in the diffuse structures, and comment on their variation with respect to diffusion time.

(6). Consider the scattering of electrons from the second to the first miniband in a superlattice, which consists of alternating layers of two different $Ga_{1-x}Al_xAs$ alloys. Discuss how each of the scattering processes considered in this chapter could be maximised by adjusting the materials used in the structure, and the operating conditions.

References

[1] J. H. Davies, *The Physics of Low-dimensional Semiconductors: An Introduction*, Cambridge University Press, Cambridge, 1997.

[2] R. Loudon, *The Quantum Theory of Light*, Oxford University Press, Oxford, Second edition, 1983.

[3] P. S. Zory, *Quantum Well Lasers*, Academic Press, Boston, 1993.

[4] M. J. Kelly, *Low Dimensional Semiconductors: Materials, Physics, Technology, Devices*, Clarendon Press, Oxford, 1995.

[5] F. W. Sears and G. L. Salinger, *Thermodynamics, Kinetic Theory and Statistical Thermodynamics*, Addison-Wesley, Reading, MA, Third edition, 1975.

[6] N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Saunders College Publishing, Philadelphia, 1976.

[7] B. K. Ridley, *Quantum Processes in Semiconductors*, Clarendon Press, Oxford, Third edition, 1993.

[8] M. Lundstrom, *Fundamentals of Carrier Transport*, Modular Series on Solid State Devices. Addison-Wesley, Reading, MA, 1990.

[9] B. K. Ridley, *Electrons and Phonons in Semiconductor Multilayers*, Cambridge University Press, Cambridge, 1997.

[10] S. Adachi, *GaAs and Related Materials: Bulk Semiconducting and Superlattice Properties*, World Scientific, Singapore, 1994.

[11] T. Piorek, *Aspects of low-dimensional diluted semimagnetic structures*, Ph.D. thesis, University of Hull, 1996.

[12] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover Publications, New York, 1965.

[13] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, London, Sixth edition, 2000.

[14] S.-H. Park, D. Ahn, and Y.-T. Lee, 'Screening effects on electron–longitudinal optical-phonon intersubband scattering in wide quantum well and comparison with experiment', *Jpn. J. Appl. Phys.*, **39**(12R):6601, 2000.

[15] B. K. Ridley, 'Electron–hybridon interaction in a quantum well', *Phys. Rev. B*, **47**(8):4592–4602, 1993.

[16] E. Molinari, C. Bungaro, M. Gulia, P. Lugli, and H. Rücker, 'Electron–phonon interactions in two-dimensional systems: a microscopic approach', *Semicond. Sci. Technol.*, **7**:B67, 1992.

[17] P. Kinsler, R. W. Kelsall, and P. Harrison, 'Interface phonons in asymmetric quantum well structures', *Superlatt. Microstruct.*, **25**:163–166, 1999.

[18] R. L. Liboff, *Introductory Quantum Mechanics*, Addison–Wesley, San Francisco, Fourth edition, 2003.

[19] P. Kinsler, P. Harrison, and R. W. Kelsall, 'Intersubband electron–electron scattering in asymmetric quantum wells designed for far-infrared emission', *Phys. Rev. B*, **58**(8):4771–4778, 1998.

[20] S. M. Goodnick and P. Lugli, 'Effect of electron–electron scattering on non-equilibrium transport in quantum well systems', *Phys. Rev. B*, **37**:2578, 1988.

[21] J. H. Smet, C. G. Fonstad, and Q. Hu, 'Intrawell and interwell intersubband transitions in multiple quantum wells for far-infrared sources', *J. Appl. Phys.*, **79**(12):9305–9320, 1996.

[22] J. M. Ziman, *Electrons and Phonons*, Oxford University Press, Oxford, 1960.

[23] N. Takenaka, M. Inoue, and Y. Inuishi, 'Influence of inter-carrier scattering on hot electron distribution function in GaAs', *J. Phys. Soc. Japan*, **47**:861, 1979.

[24] M. Moško, A. Mošková, and V. Cambel, 'Carrier–carrier scattering in photoexcited intrinsic GaAs quantum wells and its effect on femtosecond plasma thermalization', *Phys. Rev. B*, **51**(23):16860–16866, 1995.

[25] P. Harrison and R. W. Kelsall, 'The relative importance of electron–electron and electron–phonon scattering in terahertz quantum cascade lasers', *Solid State Electr.*, **42**:1449–1451, 1998.

[26] C. Sirtori, J. Faist, F. Capasso, D. L. Sivco, A. L. Hutchinson, and A. Y. Cho, 'Long wavelength infrared ($\lambda \approx 11\ \mu$m) quantum cascade lasers', *Appl. Phys. Lett.*, **69**:2810, 1996.

[27] P. F. Maldague, 'Many-body corrections to the polarizability of the two-dimensional electron gas', *Surf. Sci.*, **73**:296–302, 1978.

[28] T. Ando, A. B. Fowler, and F. Stern, 'Electronic properties of two-dimensional systems', *Rev. Mod. Phys.*, **54**(2):437–672, 1982.

[29] P. C. Slingerland, *Temperature effects and transport phenomena in terahertz quantum cascade lasers*, Ph.D. thesis, University of Massachusetts Lowell, 2011.

[30] P. Slingerland, C. Baird, and R. H. Giles, 'Application of multi-subband self-consistent energy balance method to terahertz quantum cascade lasers', *Semicond. Sci. Technol.*, **27**(6):065009, 2012.

[31] P. Kinsler, P. Harrison, and R. W. Kelsall, 'Intersubband terahertz lasers using four-level asymmetric quantum wells', *J. Appl. Phys.*, **85**:23–28, 1999.

[32] K. Donovan, P. Harrison, and R. W. Kelsall, 'Stark ladders as tunable far-infrared emitters', *J. Appl. Phys.*, **84**:5175–5179, 1998.

[33] M. Dür, S. M. Goodnick, and P. Lugli, 'Monte Carlo simulation of intersubband relaxation in wide, uniformly doped GaAs/Al_xGa_{1-x}As quantum wells', *Phys. Rev. B*, **54**(24):17794–17804, 1996.

[34] F. Stern and W. E. Howard, 'Properties of semiconductor surface inversion layers in the electric quantum limit', *Phys. Rev.*, **163**(3):816–835, 1967.

[35] G. Bastard, *Wave Mechanics Applied to Semiconductor Heterostructures*, Monographies de physique. Halsted Press, New York, 1988.

[36] M. J. Kearney and A. I. Horrell, 'The effect of alloy scattering on the mobility of holes in a quantum well', *Semicond. Sci. Technol.*, **13**(2):174, 1998.

[37] K. S. Yoon, G. B. Stringfellow, and R. J. Huber, 'Monte Carlo calculation of velocity-field characteristics in GaInAs/InP and GaInAs/AlInAs single-well heterostructures', *J. Appl. Phys.*, **62**(5):1931–1936, 1987.

[38] J. A. Van Vechten and T. K. Bergstresser, 'Electronic structures of semiconductor alloys', *Phys. Rev. B*, **1**(8):3351–3358, 1970.

[39] H. Sakaki, T. Noda, K. Hirakawa, M. Tanaka, and T. Matsusue, 'Interface roughness scattering in GaAs/AlAs quantum wells', *Appl. Phys. Lett.*, **51**(23):1934–1936, 1987.

[40] S. Tsujino, A. Borak, E. Müller, M. Scheinert, C. V. Falub, H. Sigg, D. Grützmacher, M. Giovannini, and J. Faist, 'Interface-roughness-induced broadening of intersubband electroluminescence in p-SiGe and n-GaInAs/AlInAs quantum-cascade structures', *Appl. Phys. Lett.*, **86**(6):062113, 2005.

[41] T. Unuma, M. Yoshita, T. Noda, H. Sakaki, and H. Akiyama, 'Intersubband absorption linewidth in GaAs quantum wells due to scattering by interface roughness, phonons, alloy disorder, and impurities', *J. Appl. Phys.*, **93**(3):1586–1597, 2003.

[42] M. Califano, N. Q. Vinh, P. J. Phillips, Z. Ikonić, R. W. Kelsall, P. Harrison, C. R. Pidgeon, B. N. Murdin, D. J. Paul, P. Townsend, J. Zhang, I. M. Ross, and A. G. Cullis, 'Interwell relaxation times in p-Si/SiGe asymmetric quantum well structures: Role of interface roughness', *Phys. Rev. B*, **75**(4):045338, 2007.

[43] A. Valavanis, Z. Ikonić, and R. W. Kelsall, 'Intersubband carrier scattering in n- and p-Si/SiGe quantum wells with diffuse interfaces', *Phys. Rev. B*, **77**(7):075312, 2008.

[44] U. Bockelmann and G. Bastard, 'Phonon scattering and energy relaxation in two, one and zero-dimensional electron gases', *Phys. Rev. B*, **42**:8947, 1990.

[45] L. Zheng and S. Das Sarma, 'Inelastic lifetimes of confined two-component electron systems in semiconductor quantum-wire and quantum-well structures', *Phys. Rev. B*, **54**(19):13908–13914, 1996.

[46] E. H. Hwang and S. Das Sarma, 'Electron–phonon and electron–electron interactions in one-dimensional GaAs quantum wire nanostructures', *Superlatt. Microstruct.*, **21**:1, 1997.

[47] C. R. Bennett and B. Tanatar, 'Energy relaxation via confined and interface phonons in quantum-wire systems', *Phys. Rev. B*, **55**:7165, 1997.

[48] M. Brasken, M. Lindberg, and J. Tulkki, 'Carrier dynamics in strain-induced quantum dots', *phys. stat. sol. (a)*, **164**:427, 1997.

[49] K. Kral and Z. Khas, 'Electron self-energy in quantum dots', *Phys. Rev. B*, **57**:R2061, 1998.

[50] P. K. Basu, *Theory of Optical Processes in Semiconductors*, Clarendon Press, Oxford, 1997.

[51] J. S. Blakemore, 'Semiconducting and other major properties of gallium arsenide', *J. Appl. Phys.*, **53**(10):R123–R181, 1982.

11

Optical properties of quantum wells

with D. Indjin

University of Leeds

11.1 Carrier–photon scattering

The emission (or absorption) of light by a charge carrier, whether an electron or a hole, is essentially a scattering event between an initial state i and a final state f, and as such can be understood using the methods described in Chapter 10. In this case, the electromagnetic field is the time-dependent perturbation $\tilde{\mathcal{H}}$ which induces the scattering event [1], and the transition rate from the initial electronic state $|i\rangle$ to the final state $|f\rangle$ is given by Fermi's golden rule (equation (10.19)). The photon momentum is assumed to be zero, and hence momentum conservation does not have to be considered with an additional δ-function, with the physical implication of this being that transitions on the band diagrams are always vertical whether they are between bands (interband) or within the same band (intraband or intersubband) (see Fig. 11.1).

Although a great deal has been achieved within the envelope function approximation, it is necessary now to reintroduce the rapidly varying component of the wave function, writing this as the product $u\psi$ (see Section 1.7). Then, as reiterated by Adachi ([2], p. 476) and Basu ([3], p. 299), Coon and Karunasiri [4] showed that the matrix element in equation (10.19) becomes:

$$\langle f|\tilde{\mathcal{H}}|i\rangle = \langle u_f|\ \tilde{\mathcal{H}}\ |u_i\rangle_{\text{cell}} \langle\psi_f|\psi_i\rangle + \langle u_f|u_i\rangle_{\text{cell}} \langle\psi_f|\ \tilde{\mathcal{H}}\ |\psi_i\rangle \tag{11.1}$$

In the case of interband transitions between the conduction and valence band, the second term is zero since the Bloch functions u_f and u_i at the same point in the Brillouin zone, in two different bands, are orthogonal, i.e.

$$\langle u_f|u_i\rangle_{\text{cell}} = \int_{\text{cell}} u_f(\mathbf{r})u_i(\mathbf{r})\ \ \mathrm{d}\mathbf{r} = 0 \tag{11.2}$$

and therefore:

$$\langle f|\tilde{\mathcal{H}}|i\rangle = \langle u_f|\ \tilde{\mathcal{H}}\ |u_i\rangle_{\text{cell}} \langle\psi_f|\psi_i\rangle \tag{11.3}$$

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

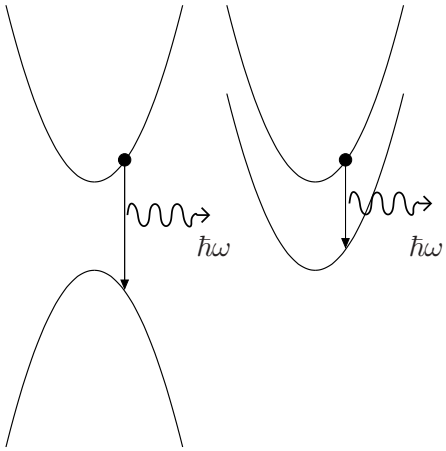


Figure 11.1: Carrier–photon scattering events result in vertical transitions: interband (left) and intraband (intersubband) (right)

Hence the envelope function overlap integral, $\langle\psi_f|\psi_i\rangle$, determines which transitions are allowed and which are forbidden, which was the result quoted earlier in Section 6.8 during the discussion of excitons.

Radiative interband transitions are, of course, very important in both light sources (see, for example, [5]) and detectors (see, for example, [6, 7, 8]). This mechanism gives access to a frequency range from the near-infrared [9] to the blue, depending on the material system employed. For a review of wide-bandgap III–V emitters and lasers, see Nakamura [10]; for II–VI, see Gunshor and Nurmikko [11]. Much discussion has already centred around interband transitions in the calculation of energy levels (see Chapters 2 and 3), exciton binding energies, and oscillator strengths (see Chapter 6); for further information, see Ivchenko and Pikus ([12], p. 162), or Bastard ([13], p. 237).

For intersubband transitions the first term on the right-hand side of equation (11.1) is zero, and since the subband envelope functions ψ_f and ψ_i are both eigenfunctions of the same Hermitian operator (the conduction- or valence-band Hamiltonian), they are therefore orthogonal, i.e.

$$\langle\psi_f|\psi_i\rangle = \int_{\text{all space}} \psi_f(\mathbf{r})\psi_i(\mathbf{r})\ \mathrm{d}\mathbf{r} = 0 \tag{11.4}$$

Therefore, the matrix element becomes:

$$\langle f|\tilde{\mathcal{H}}|i\rangle = \langle u_f|u_i\rangle_{\text{cell}} \langle\psi_f|\ \tilde{\mathcal{H}}\ |\psi_i\rangle \tag{11.5}$$

where the time-dependent perturbing potential $\tilde{\mathcal{H}}$ is given by:

$$\tilde{\mathcal{H}} = \frac{e}{m^*}\mathbf{A}_\bullet\mathbf{p} \tag{11.6}$$

(see [2], p. 476, [14] and, most appropriately, [15], p. 184). The vector $\mathbf{A}$ is the vector potential of the electromagnetic field and hence consists of a magnitude and a direction, with

the latter being represented by the unit polarisation vector $\hat{\mathbf{e}}$. The time-dependent magnitude is, without loss of generality, simply given as:

$$A(t) = A_0 \cos(\omega t) \tag{11.7}$$

where ω is the angular frequency of the electromagnetic radiation (photon energy is $\hbar\omega$).

Substituting equation (11.6) and the linear momentum operator, $\mathbf{p} = -i\hbar\nabla$, into the scattering matrix element gives:

$$\langle f|\tilde{\mathcal{H}}|i\rangle = \frac{e}{m^*} P_{if} A(t) \tag{11.8}$$

where

$$P_{if} = -i\hbar\langle f|\hat{\mathbf{e}}_\bullet \nabla|i\rangle \tag{11.9}$$

Substituting in the complete form of the wave functions in a two-dimensional system gives:

$$P_{if} = -i\frac{\hbar}{A}\iiint \psi_f^*(z)\mathrm{e}^{i\mathbf{k}_f\bullet\mathbf{r}}\,\hat{\mathbf{e}}_\bullet\nabla\psi_i(z)\mathrm{e}^{-i\mathbf{k}_i\bullet\mathbf{r}}\,\mathrm{d}z\,\mathrm{d}\mathbf{r} \tag{11.10}$$

In this case of intersubband scattering, the envelope wave functions ψ_i and ψ_f are functions of the displacement along the growth (z-)axis only. Therefore, the gradient operator may be expanded as:

$$\nabla\psi_i(z)\mathrm{e}^{-i\mathbf{k}_i\bullet\mathbf{r}} = \mathrm{e}^{-i\mathbf{k}_i\bullet\mathbf{r}}\left[-\hat{\mathbf{i}}(ik_x\psi_i(z)) - \hat{\mathbf{j}}(ik_y\psi_i(z)) + \hat{\mathbf{k}}\frac{\partial\psi_i}{\partial z}\right] \tag{11.11}$$

Now, by recalling that the initial and final wave functions are orthonormal, i.e. $\langle f|i\rangle = 0$ for a pair of non-identical states, it can be shown that the components of P_{if} in the x- and y-directions are zero. As such, the matrix element is reduced to:

$$P_{if} = -i\frac{\hat{e}_z\hbar}{A}\iiint \psi_f^*(z)\mathrm{e}^{i\mathbf{k}_f\bullet\mathbf{r}}\,\frac{\partial\psi_i}{\partial z}\mathrm{e}^{-i\mathbf{k}_i\bullet\mathbf{r}}\,\mathrm{d}z\,\mathrm{d}\mathbf{r} \tag{11.12}$$

This implies that transitions are only allowed when there is a component of the polarisation vector $\hat{\mathbf{e}}$ along the growth (z-)axis [16, 17], which means that no intersubband absorption occurs for normal (i.e. along the growth (z-)axis) incident light. This is illustrated in the top diagram of Fig. 11.2. For the case of normal-incidence light, there is no component of the polarisation vector $\hat{\mathbf{e}}$ along the growth (z-)axis of the heterostructure, and hence no intersubband absorption occurs. When the incident light is at an angle, the z-component $\hat{e}_z$ is non-zero, and some absorption is allowed. This is a major difference between intersubband and interband transitions.

In addition, any laser made from intersubband transitions, whether optically pumped or electrically injected, as illustrated schematically in the bottom diagram of Fig. 11.2, will be an edge emitter [18, 19].

These conclusions about natural device geometries are valid for the simple single parabolic band wave functions considered in the first instance. However, there is a wealth of literature available about possible band structure effects, including non-parabolicity, anisotropy, and degeneracy which could break these selection rules (for experimental work, see, for

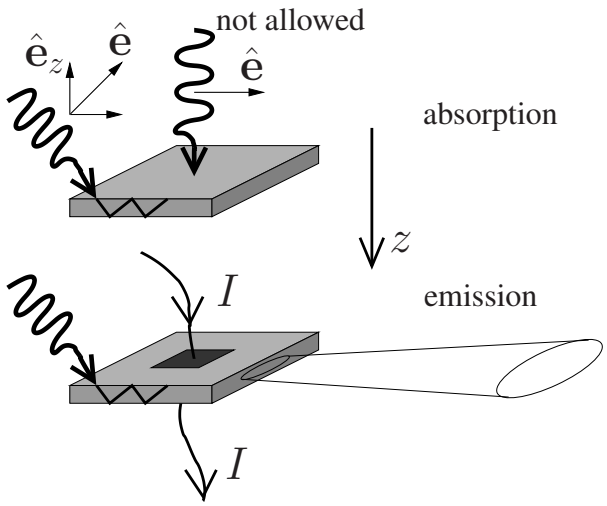


Figure 11.2: Natural intersubband device geometries

example, [20], and for a detailed theoretical analysis, see Batty and Shore [21] and references therein).

Returning to equation (11.12), the integrals can now be rearranged as:

$$P_{if} = -i\frac{\hat{e}_z\hbar}{A}\iint \mathrm{e}^{i(\mathbf{k}_f-\mathbf{k}_i)\bullet\mathbf{r}}\,\mathrm{d}\mathbf{r}\int \psi_f^*(z)\,\frac{\partial\psi_i}{\partial z}\,\mathrm{d}z \tag{11.13}$$

The integral over $\mathbf{r}$ evaluates to the in-plane area, A, giving the simplified result:

$$P_{if} = -i\hbar\,\hat{e}_z\int \psi_f^*(z)\,\frac{\partial\psi_i}{\partial z}\,\mathrm{d}z = \hat{e}_z\int \psi_f^*(z)\hat{p}_z\psi_i(z)\mathrm{d}z \tag{11.14}$$

i.e. the momentum matrix element in the z-direction multiplied by the polarisation component of the radiation in the same direction.

This expression may be simplified further by making use of the well-known commutation relation between the non-perturbed Hamiltonian and the position operator:

$$\hat{p}_z = \frac{m^*}{i\hbar}[\mathcal{H}_0, z] = \frac{m^*}{i\hbar}(\mathcal{H}_0 z - z\mathcal{H}_0) \tag{11.15}$$

Substituting this into equation (11.14) gives:

$$P_{if} = \hat{e}_z\frac{m^*}{i\hbar}\int \psi_f^*(z)(\mathcal{H}_0 z - z\mathcal{H}_0)\psi_i(z)\mathrm{d}z = -i\frac{m^*}{\hbar}(E_f - E_i)\int_z \psi_f^*(z)z\psi_i(z)\,\mathrm{d}z \tag{11.16}$$

The final expression on the right-hand side contains the dipole matrix element, which is simpler to calculate than the momentum matrix element, giving the final form $P_{if} = -i\hat{e}_z m^*\omega d_{if}$.

Note that the derivation here has assumed a constant effective mass with respect to energy (i.e. parabolic bands). In the case of non-parabolic subbands, however, the wave functions are *not* orthogonal to each other since the Schrödinger equation is a nonlinear differential equation, and the energy-dependent mass must be accounted for when computing the transition matrix element. Although this is beyond the scope of the present discussion, a derivation is presented in [22].

This form of the momentum matrix element can now be used to find the electron–photon scattering rate at a given time, in response to the oscillating light field. Substituting equation (11.8) into (10.19) gives:

$$\mathrm{w}_{if}(t) = \frac{2\pi}{\hbar}\left(\frac{e}{m^*}\right)^2 |P_{if}|^2 A(t)^2 \delta(E_f - E_i \pm \hbar\omega) \tag{11.17}$$

Now, the time-independent scattering rate w_{if} is found by averaging over the period of the incident radiation $T_\mathrm{A} = 2\pi/\omega$:

$$\mathrm{w}_{if} = \frac{2\pi}{\hbar}\left(\frac{e}{m^*}\right)^2 |P_{if}|^2 \frac{A_0^2}{2} \delta(E_f - E_i \pm \hbar\omega) \tag{11.18}$$

where the time-dependent term $A(t)$ (equation (11.7)) is averaged as:

$$\overline{A(t)^2} = \frac{1}{T_\mathrm{A}} \int_0^{T_\mathrm{A}} A_0^2 \cos^2(\omega t)\, \mathrm{d}t = \frac{A_0^2}{2} \tag{11.19}$$

The *net* scattering rate can be found by summing over all possible interaction events of photons being absorbed or emitted in electronic transitions between the initial state $\langle\, i, \mathbf{k}_i|$ and the final state $\langle f, \mathbf{k}_f|$, i.e.

$$\begin{aligned} W_{if} &= \sum_{k_x}\sum_{k_y} \mathrm{w}_{if}[\underbrace{f^{\mathrm{FD}}(E_i)(1 - f^{\mathrm{FD}}(E_f))}_{\text{absorption term}} - \underbrace{f^{\mathrm{FD}}(E_f)(1 - f^{\mathrm{FD}}(E_i))}_{\text{emission term}}] = \\ &= \sum_{k_x}\sum_{k_y} \mathrm{w}_{if}[f^{\mathrm{FD}}(E_i) - f^{\mathrm{FD}}(E_f)] \end{aligned} \tag{11.20}$$

Generally speaking, in a two-dimensional system, the effective mass is z-dependent ($m^* = m(z)$), and strictly the Schrödinger equation should be solved for each in-plane wave vector k_{xy}. In order to simplify the procedure, it is a reasonable approximation, based on time-independent perturbation theory (see, for example, [23]), to introduce an *in-plane effective mass* in which case the Schrödinger equation can be solved only for $k_{xy} = 0$ and then for each state $|j\rangle$ (where j is either the initial or final state, i.e. $j = i, f$) calculate the in-plane effective mass. Therefore, the eigenenergy corresponding to an arbitrary in-plane wave vector k_{xy} is given as a small perturbation of the energy obtained by solving the Schrödinger equation for $k_{xy} = 0$, i.e.

$$E_j(k_{xy}^2) = E_j(k_{xy}^2 = 0) + \int \psi_{j_0}^*(z) \frac{\hbar^2 k_{xy}^2}{2m(z)} \psi_{j_0}(z)\, \mathrm{d}z = E_j(0) + \frac{\hbar^2 k_{xy}^2}{2m_{\mathrm{t}j}} \tag{11.21}$$

hence, the in-plane effective mass is calculated as:

$$\frac{1}{m_{\mathrm{t}j}} = \langle j|\, 1/m(z)\, |j\rangle = \int \frac{1}{m(z)} |\psi_{j_0}(z)|^2 \, \mathrm{d}z \tag{11.22}$$

where the wave function $\psi_{j_0}(z)$ is also obtained as a solution of the Schrödinger equation for $k_{xy} = 0$. Obviously, introduction of the in-plane effective mass is an approximation analogous to the parabolic band approximation.

11.2 Spontaneous emission lifetime

The lifetime for intersubband spontaneous radiative emission is obtained by summing equation (11.20) over all photon modes. As will be discussed in Section 11.3, this requires some assumption on the cavity, i.e. the region of dielectric that the modes occupy. Conveniently, Smet *et al*. [14] quote the results as:

$$\frac{1}{\tau_i} = W_{if} = \frac{e^2 n \omega^2}{6\pi\epsilon m^* c^3} O_{if} \tag{11.23}$$

where n is the refractive index at the emission wavelength, for a three-dimensional distribution of photon modes, and:

$$W_{if} = \frac{e^2 \omega}{4\epsilon m^* c^2 W_z} O_{if} \tag{11.24}$$

for a two-dimensional photonic density of states. The latter is a common scenario in a quantum well system which is often surrounded by barriers and/or a substrate of different dielectric material, which can act as a microcavity producing confinement of the optical modes along the growth (z-)axis, as illustrated in Fig. 11.3.

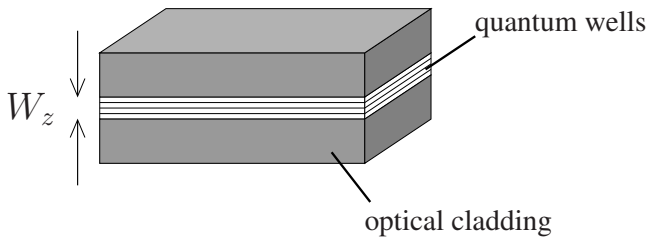


Figure 11.3: Schematic representation of a typical semiconductor quantum well system surrounded by dielectric material, thus producing confinement of optical modes

The oscillator strength O_{if} is dependent upon the dipole matrix element [18], i.e.

$$O_{if} = \frac{2m^*\omega}{\hbar} \left|\langle \psi_i \,|z|\, \psi_f \rangle\right|^2 \tag{11.25}$$

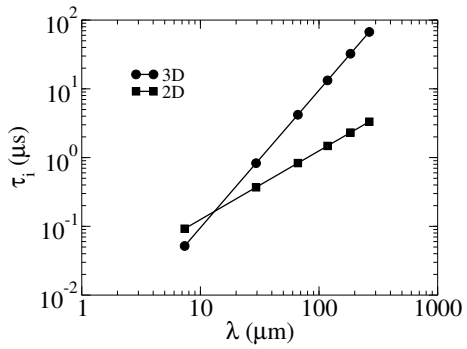


Figure 11.4: Comparison of the three-dimensional (3D) and two-dimensional (2D) spontaneous radiative emission lifetimes

Given that the matrix element is squared, then it is equivalent (see [24], p. 26), to that used earlier to derive the selection rules.

Figure 11.4 compares the results of calculations of the three-dimensional and two-dimensional spontaneous radiative emission lifetimes, between the two lowest conduction-band eigenstates of a GaAs infinitely deep quantum well; the emission wavelength has been varied by adjusting the quantum well width. This figure shows that for short-wavelength emission in the mid-infrared (less than $10\,\mu\mathrm{m}$), the radiative lifetime calculated by using the 3D distribution of photon modes is shorter than that calculated with the 2D distribution, with W_z taken as $3\,\mu\mathrm{m}$. However, as the wavelength increases, and the emission energy decreases, shorter lifetimes for radiative emission are obtained with the 2D equation. Figure 11.4 basically confirms the work of Smet *et al.* [14]. The refractive index n was calculated by using the first-order Sellmeier equation [25] and the data of Seraphin and Bennett [26].

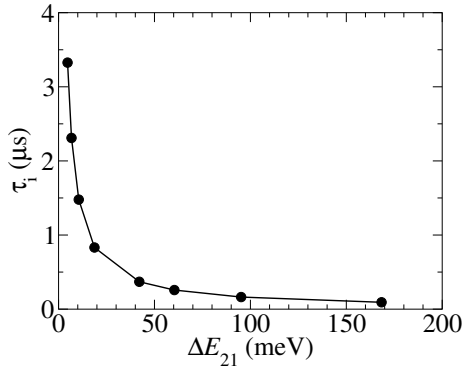


Figure 11.5: The 2D spontaneous lifetime as a function of subband separation

Figure 11.5 plots the 2D spontaneous radiative lifetime as in Fig. 11.4, but this time as a function of the subband separation. It can be seen that the lifetime increases as the energy separation decreases, with what appears to be a hyperbolic dependence. This is the reason why intersubband emitters and lasers are thought to be more difficult to fabricate at longer wavelengths (see, for example, [27]). Figure 11.6 confirms the hyperbolic dependence of the

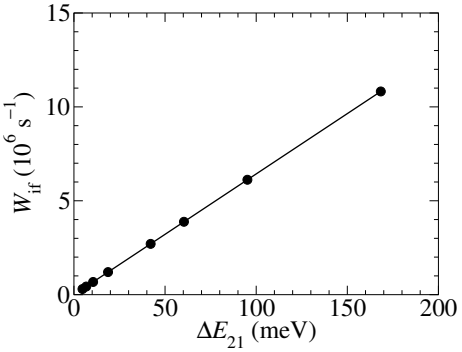


Figure 11.6: The 2D spontaneous emission rate as a function of subband separation

lifetime on the subband separation by plotting its inverse, i.e. the scattering rate, against $\Delta E_{21} = E_2 - E_1$. Notice the striking linearity implying that

$$W_{if} \propto \Delta E_{21} \tag{11.26}$$

This is not as obvious as it seems, for in the expression for this scattering rate, the ω in the numerator cancels with the ω in the denominator. Hence, the functional dependencies are controlled by the matrix element.

The temperature independence of the spontaneous radiative lifetime *in this model* should also be noted. In fact, this is mainly because for simplicity the material dependencies on temperature have been ignored, in order to deduce as general results as are possible. In reality, the energy gaps do change slightly and hence the emission wavelength also changes (see Adachi [2] for further information). While in these models the emission frequency does not change, the intensity, i.e. the number of photons emitted, does change—in fact, the latter decreases. This is because of the increased competition with the non-radiative channels.

11.3 Intersubband absorption in quantum wells

The previous section discussed the *rate* of spontaneous emission of photons in a quantum well system. In the case of absorption of light, it is usually more useful to determine an absorption coefficient, i.e. a measure of how much of the incident radiation is absorbed within the structure. The average intensity of incident radiation is by definition given as:

$$I_{\mathrm{p}_0} = w_{\mathrm{em}} v_\Phi = w_{\mathrm{em}} \frac{c}{\bar{n}} \tag{11.27}$$

where v_Φ is the velocity of light in the medium and $\bar{n}$ is the average refractive index. $w_{\rm em}$ is the density of electromagnetic energy, averaged over the time period, given as the sum of the electrostatic and magnetic energy, i.e.

$$w_{\rm em} = w_{\rm e} + w_{\rm m} = 2w_{\rm e} = 2 \times \frac{1}{2}\epsilon_{\rm r}\epsilon_0\overline{|\mathbf{E}|^2} \tag{11.28}$$

where **E** is the electric field vector calculated as:

$$\mathbf{E} = -\frac{\partial \mathbf{A}}{\partial t} = -\frac{\partial (A(t)_\bullet \mathbf{e})}{\partial t} = \omega A_0 \sin(\omega t)_\bullet \mathbf{e} \tag{11.29}$$

Substituting the mean-squared value of equation (11.29) into equation (11.28) gives:

$$w_{\rm em} = \frac{1}{2}\epsilon_{\rm r}\epsilon_0\omega^2 A_0^2 \tag{11.30}$$

Thus, equation (11.27), together with $\epsilon_{\rm r} = \bar{n}^2$, gives an expression for the average incident intensity of radiation as:

$$I_{\rm p_0} = \frac{1}{2}\epsilon_0 \bar{n} c \omega^2 A_0^2 \tag{11.31}$$

On the other hand, the intensity of light at a depth z in the semiconductor structure is given by the Beer–Lambert absorption function:

$$I_{\rm p}(z) = I_{\rm p_0} \exp(-\alpha_{if} z) \tag{11.32}$$

which, assuming a small structure and/or small absorption ($\alpha_{if} z \ll 1$) reads:

$$I_{\rm p}(z) = I_{\rm p_0}(1 - \alpha_{if} z) \tag{11.33}$$

The above equation defines the absorption coefficient in the form:

$$\alpha_{if} = \frac{I_{\rm p_0} - I_{\rm p}(z)}{z I_{\rm p_0}} = \frac{W_{if}\hbar\omega}{V_z I_{\rm p_0}} \tag{11.34}$$

where $(I_{\rm p_0} - I_{\rm p}(z))/z = W_{if}\hbar\omega/V_z$ represents the density of electromagnetic energy absorbed in the volume $V_z = zL_xL_y$ where z is the thickness of the quantum well structure along the growth axis and L_x and L_y are the dimensions of the structure in the x- and y-directions, respectively.

Considering the whole quantum well structure shown in Fig. 11.3, with dimensions L_x, L_y and $z \to W_z$, the absorption coefficient reads:

$$\alpha_{if} = \frac{2\pi e^2}{\bar{n}\epsilon_0 \omega m_0^2 c L_x L_y W_z} \sum_{k_x} \sum_{k_y} |P_{if}|^2 F_{if} \delta(E_f - E_i - \hbar\omega) \frac{\Delta k_x}{\Delta k_x} \frac{\Delta k_y}{\Delta k_y} \tag{11.35}$$

where $F_{if} = f^{\rm FD}(E_i) - f^{\rm FD}(E_f)$ represents the difference in the populations of states $|i\rangle$ and $|f\rangle$ (the difference of Fermi–Dirac functions for the initial and final states). The factor 2 was introduced to account for electrons of both spin orientations. In the case of an in-plane finite structure, the quasi-discrete steps Δk_x and Δk_y can be introduced in the derivation,

representing the 'distance' of subsequent values of the **k** vector components k_x and k_y in the lateral dimensions. Based on the 'box' boundary conditions applied across the structure (vanishing of the wave functions at the structure edges), $\Delta k_x = \pi/L_x$ and $\Delta k_y = \pi/L_y$ and, assuming that the structure is close to being infinite ($L_x, L_y \to \infty$), $\Delta k_x \to 0$ and $\Delta k_y \to 0$ can be substituted with $\mathrm{d}k_x$ and $\mathrm{d}k_y$ and hence, equation (11.35) becomes an integral:

$$\alpha_{if} = \frac{2e^2}{\bar{n}\epsilon_0 \omega m_0^2 c \pi W_z} \int_{k_x} \int_{k_y} |P_{if}|^2 F_{if} \delta(E_f - E_i - \hbar\omega)\, \mathrm{d}k_x \mathrm{d}k_y \tag{11.36}$$

Because of the fact that, if $W_z \to +\infty$, then also $\alpha_{if} \to +\infty$, it is more convenient to represent the absorption properties of the quantum well structure in terms of the fractional absorption defined as:

$$A_{if} = \alpha_{if} W_z \tag{11.37}$$

The fractional absorption is a dimensionless quantity and is commonly given as a percentage, defining the proportion of radiation absorbed by the quantum well structure in one pass.

As the states' dispersion is axially isotropic, introducing an in-plane wave vector $k_{xy}^2 = k_x^2 + k_y^2$ transforms the integral $\int_0^\infty \int_0^\infty (\ldots)\, \mathrm{d}k_x \mathrm{d}k_y = (\pi/4) \int_0^\infty (\ldots)\, \mathrm{d}k_{xy}^2$, based on which equation (11.36) reads:

$$A_{if} = \frac{e^2}{2\bar{n}\epsilon_0 \omega m_0^2 c} \int_0^\infty |P_{if}|^2 F_{if} \delta(E_f - E_i - \hbar\omega)\, \mathrm{d}k_{xy}^2 \tag{11.38}$$

The integral can now be simplified by using the property of the δ-function that:

$$\int_{-\infty}^{+\infty} f(x)\delta(g(x)) = \sum_i \frac{f(x_\mathrm{i})}{|g'(x_\mathrm{i})|}, \quad g(x_i) = 0 \tag{11.39}$$

Here, k_{xy0} is the in-plane wave vector that is required to conserve energy, which can be obtained by solving:

$$E_f(k_{xy0}^2) - E_i(k_{xy0}^2) = \hbar\omega \tag{11.40}$$

This gives the final expression:

$$A_{if} = \frac{e^2}{2\bar{n}\epsilon_0 \omega m_0^2 c} \left[\left|P_{if}\right|^2 F_{if} \left| \frac{\mathrm{d}(k_{xy}^2)}{\mathrm{d}(E_f - E_i)} \right| \right]_{k_{xy}^2 = k_{xy0}^2} \tag{11.41}$$

In the derivation the **k**-selection rule (in-plane wave-vector conservation) was implicitly used, i.e. ($\mathbf{k}_i = \mathbf{k}_f$).

Equation (11.41) can be solved to find the fractional absorption within a structure at a range of photon frequencies, $A_{if}(\hbar\omega)$, or in other words, the *absorption spectrum* for the system.

11.4 Bound–bound transitions

The energy difference between two bound states in different subbands at in-plane wave vector k_{xy} is:

$$E_f(k_{xy0}^2) - E_i(k_{xy0}^2) = \hbar\omega_0 - \frac{\hbar^2 k_{xy0}^2}{2}\left(\frac{1}{m_{\mathrm{t}i}} - \frac{1}{m_{\mathrm{t}f}}\right) \tag{11.42}$$

where $\hbar\omega_0 = E_f(0) - E_i(0)$ is the separation between the states at the subband minima. The second term on the right-hand side accounts for the in-plane dispersion of the subbands, using the in-plane effective masses $m_{\mathrm{t}i}$ and $m_{\mathrm{t}f}$ given by equation (11.22).

This expression can be simplified further by defining an in-plane effective mass *for the transition*:

$$\frac{1}{m_{if}} = \frac{1}{m_{\mathrm{t}i}} - \frac{1}{m_{\mathrm{t}f}} \tag{11.43}$$

such that equations (11.42) and (11.40) become:

$$\hbar\omega = \hbar\omega_0 - \frac{\hbar^2 k_{xy0}^2}{2m_{if}} \tag{11.44}$$

In a system with parabolic energy bands, the effective mass is identical in each subband, and, as such, $\hbar\omega = \hbar\omega_0$ at all in-plane wave vectors. Solution of equation (11.41) will then yield a non-zero absorption at only that precise energy, and the spectrum will take the shape of a Dirac δ-function (i.e. a perfectly sharp absorption feature).

However, in reality, absorption spectra consist of broadened peaks, spread over a range of energies. For example, in non-parabolic energy bands, the effective mass typically increases with respect to energy, as described in Section 1.8. As such, $m_{\mathrm{t}f} > m_{\mathrm{t}i}$ and equation (11.44) has a solution for a range of energies $\hbar\omega < \hbar\omega_0$, with the energy decreasing as the in-plane wave vector increases. The calculated absorption spectrum will then spread over this broadened energy range, with the shape of the peak being determined by the form of the band dispersion and the distribution of carrier energies. However, in addition to band dispersion effects, numerous mechanisms contribute to intersubband absorption line broadening such as scattering processes, and inhomogeneity in the structure of devices. Therefore, a more appropriate and experimentally justified line shape is the Lorentzian given as [28]:

$$L(\hbar\omega, \hbar\omega_0) = \frac{\gamma_{ij}/2\pi}{(\hbar\omega - \hbar\omega_0)^2 + (\gamma_{ij}/2)^2}, \tag{11.45}$$

where γ_{ij} is the full width at half maximum, usually taken from experiment.

In order to incorporate the absorption line broadening, the δ-function in equation (11.41) has to be substituted with a normalised Lorentzian form $L(\hbar\omega, \hbar\omega_0)$, therefore the normalising condition $\int_{-\infty}^{\infty} L \,\mathrm{d}(\hbar\omega) = 1$ is satisfied. Hence, the expression for the intersubband absorption in quantum wells reads:

$$A_{if}(\omega) = \frac{e^2\omega}{2\overline{n}\epsilon_0 c}\,|d_{if}|^2 \int_0^{\infty} L(\hbar\omega, \hbar\omega_0) F_{if}(k_{xy}^2)\,\mathrm{d}k_{xy}^2 \tag{11.46}$$

In equation (11.21) the transition energy $\hbar\omega_0$ and the corresponding Lorentzian are independent of the in-plane wave vector (k_{xy}^2). Equation (11.46) therefore takes the form:

$$A_{if}(\omega) = \frac{e^2\omega}{2\overline{n}\epsilon_0 c}\,|d_{if}|^2\,L(\hbar\omega,\hbar\omega_0)\int_0^\infty \left(f^{\mathrm{FD}}(E_i,k_{xy}^2) - f^{\mathrm{FD}}(E_f,k_{xy}^2)\right)\,\mathrm{d}k_{xy}^2 \tag{11.47}$$

Integrating the Fermi–Dirac distribution analytically over the subband dispersion, i.e. $\int f^{\mathrm{FD}}(E_j,k_{xy}^2)\,\mathrm{d}k_{xy}^2 = 2\pi N_j$, where N_j is the sheet carrier concentration (the electron density per unit well surface) of state j, the final formula for fractional absorption on *bound–bound transitions* reads:

$$A_{if}(\omega) = \frac{e^2\omega\pi}{\overline{n}\epsilon_0 c}\,|d_{if}|^2\,L(\hbar\omega,\hbar\omega_0)\,(N_i - N_f) \tag{11.48}$$

The fractional absorption, therefore, depends on the *difference* between the populations of the initial and final states, which can be calculated using equation (2.76).

11.5 Bound–free transitions

In calculating the absorption for bound (discrete)–free (continuum) transitions one has to account for the twofold degeneracy in the continuum, and choose the two degenerate wave functions to be orthogonal, in order to avoid overcounting or undercounting the continuum states. Therefore, consideration should be given to transitions from an initial bound state to both free final states corresponding to an energy in the continuum spectrum. Bound–free transitions have practical relevance for quantum well infrared photodetectors. Introduce the notation that the index i of an initial state becomes b (bound state) and the index f of a final state becomes c (continuum state). Assuming that continuum spectrum is initially discretised, equation (11.41) reads:

$$A_{if} = \frac{4\pi\beta\omega|m_{\mathrm{bc}}|}{\bar{n}\hbar}[|d_{\mathrm{bc}}|^2 F_{\mathrm{bc}}]_{k_{xy}^2=k_{xy_0}^2} \tag{11.49}$$

where $\beta = e^2/(4\pi\epsilon_0\hbar c)$, $1/m_{\mathrm{bc}} = 1/m_{\mathrm{b}} + 1/m_{\mathrm{c}}$ and $|d_{\mathrm{bc}}|^2 = |d_{\mathrm{bc}_1}|^2 + |d_{\mathrm{bc}_2}|^2$ is the contribution of the two free states to the dipole matrix element (note that in symmetric structures because of wave-function parity, one of the two terms is zero). The energy of the free state is given as:

$$E_{\mathrm{c}}(k_{xy}^2) = \frac{\hbar^2}{2m_{\mathrm{c}}}(k_z^2 + k_{xy}^2) \tag{11.50}$$

while the energy of the bound state is given by equation (11.21) with index j becoming b. The in-plane wave vector k_{xy0} is obtained by solving equation (11.40) as:

$$k_{xy0}^2 = \frac{m_{\mathrm{b}}}{m_{\mathrm{c}} - m_{\mathrm{b}}}\left[k_z^2 - \frac{2m_{\mathrm{c}}}{\hbar^2}\left(E_{\mathrm{b}}\,(0) + \hbar\omega\right)\right] \tag{11.51}$$

The factor in square brackets should not be smaller than zero which, in turn, determines the minimal value of k_z:

$$k_{z_{\min}} = \begin{cases} 0, & \hbar\omega + E_{\mathrm{b}}(0) \le 0 \\ \sqrt{\frac{2m_{\mathrm{c}}}{\hbar^2}[\hbar\omega + E_{\mathrm{b}}(0)]}, & \hbar\omega + E_{\mathrm{b}}(0) > 0 \end{cases}$$

The total bound–continuum fractional absorption is found by a summation over the discretised continuum spectrum:

$$A_{\mathrm{bc}} = \sum_{k_z} \frac{4\pi\beta\omega|m_{\mathrm{bc}}|}{\bar{n}\hbar} F_{\mathrm{bc}} d_{\mathrm{bc}}^2 \frac{\Delta k_z}{\Delta k_z} \tag{11.52}$$

where the normalisation constant of a continuum wave function is proportional to $1/\sqrt{L_z}$, i.e. the dipole matrix element reads:

$$d_{\mathrm{bc}} = \frac{1}{\sqrt{L_z}} d_{\mathrm{bc}}^* = \frac{1}{\sqrt{L_z}} \langle \psi_{\mathrm{c}} | \, z \, | \psi_{\mathrm{b}} \rangle \tag{11.53}$$

where d_{bc}^* is the dipole matrix element calculated with the free state wave function without the normalisation factor $1/\sqrt{L_z}$. Again, the assumption can be made that the structure is close to being infinite ($L_z \to \infty$), and box boundary conditions can be applied in the z-direction (vanishing of wave functions at structure edges) to obtain $\Delta k_z = \pi/L_z$. $\Delta k_z \to 0$ can be substituted with $\mathrm{d}k_z$, and hence equation (11.53) becomes an integral and the total bound–continuum fractional absorption is found by integration over the continuum states [29]:

$$A_{\mathrm{bc}} = \frac{4\beta\omega m_{\mathrm{bc}}}{\bar{n}\hbar} \int_{k_{z_{\min}}}^{+\infty} [F_{\mathrm{bc}} d_{\mathrm{bc}}^*]_{k_{xy}^2 = k_{xy0}^2} \, \mathrm{d}k_z \tag{11.54}$$

where F_{bc} denotes the difference of Fermi–Dirac functions for bound and free states. Obviously, in the case of a symmetric structure, taking into account wave-function parity, only transitions from even bound states to odd free states or from odd bound states to even free states will exist.

It should be noted that Lorentzian broadening was not included in the above equations because in bound–free transitions the 'band structure' linewidth is much larger than the Lorentzian linewidth; therefore, in contrast to bound–bound transitions, the latter effect may be neglected in realistic structures. In addition, thanks to the another integration in the bound–free absorption calculation, the experimentally non-realistic feature of the δ-function peak is lost, hence inclusion of the Lorentzian is not necessary.

11.6 Rectangular quantum well

By way of an example, consider a symmetric square quantum well with finite barrier height V_0 and well width l_{w}, based on the $\mathrm{GaAs/Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ system. For simplicity, the conduction-band edge in bulk AlGaAs is chosen as the reference energy level. The eigenstates of the system are found quasi-analytically using the techniques in Section 2.7 and are plotted in Fig. 11.7.

Once the eigenenergies and wave functions are known, the bound–bound dipole matrix element d_{if} in equation (11.48) can be calculated straightforwardly. Figure 11.8 shows the calculated bound–bound absorption spectra as a function of the photon energy $\hbar\omega$, for a rectangular GaAs quantum well embedded in $\mathrm{Al}_{0.33}\mathrm{Ga}_{0.67}\mathrm{As}$ bulk at a temperature $T = 77\,\mathrm{K}$. The thickness of the well is $l_{\mathrm{w}} = 80\,\text{Å}$, which supports two bound states. The particular value of the Al mole fraction of $x = 0.33$ gives the depth of the potential well of

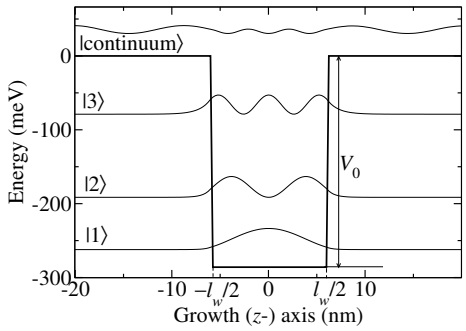


Figure 11.7: A rectangular AlGaAs quantum well with three bound states. Eigenfunctions squared are also presented as well as a representative of the continuum spectrum wave functions

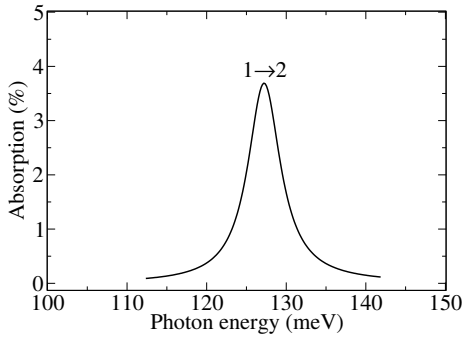


Figure 11.8: Calculated bound–bound absorption as a function of the photon energy $\hbar\omega$ for an $8\,\mathrm{nm}$ thick GaAs quantum well in $\mathrm{Al}_{0.33}\mathrm{Ga}_{0.67}\mathrm{As}$ bulk at $T = 77\,\mathrm{K}$. The value of the Fermi energy is $E_\mathrm{F} = -230\ \mathrm{meV}$. The position of the absorption peak corresponds to a difference of energies of the two bound levels in the quantum well

$V_0 = 286\,\mathrm{meV}$ and the effective mass of the barrier material $m_\mathrm{B} = 0.094m_0$. The calculated energies of the ground state and the first excited state with respect to the conduction-band edge in bulk are $E_1 = -243\ \mathrm{meV}$ and $E_2 = -116\ \mathrm{meV}$, respectively. The doping level was chosen to provide a value of the Fermi energy $E_\mathrm{F} = -230\,\mathrm{meV}$ at a given cryogenic temperature and the Lorentzian broadening (full width at half maximum) was taken to be $\gamma_{12} = 5\,\mathrm{meV}$ [30]. Figure 11.8 illustrates the characteristic bell-shaped intersubband absorption spectrum, which has a calculated magnitude of around 4% with the position of the peak corresponding to a difference of energies of the first excited and ground (bound) state $E_2 - E_1 = 127\,\mathrm{meV}$.

In the second example, a somewhat larger $\mathrm{GaAs/Al}_{0.33}\mathrm{Ga}_{0.67}\mathrm{As}$ quantum well, $160\,\text{Å}$ thick, with four bound levels is considered. The calculated energies of the bound states are $E_1 = -271\ \mathrm{meV}$, $E_2 = -226\ \mathrm{meV}$, $E_3 = -153\ \mathrm{meV}$ and $E_4 = -54\ \mathrm{meV}$, and the Fermi

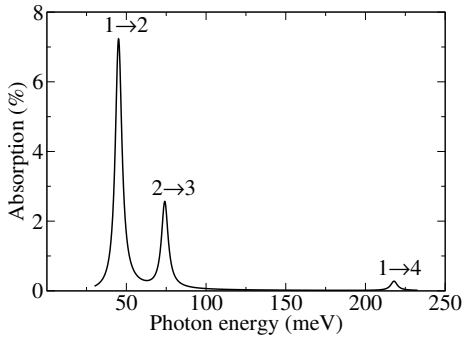


Figure 11.9: Calculated bound–bound absorption as a function of the photon energy $\hbar\omega$ for a $16\,\mathrm{nm}$ thick GaAs quantum well in $\mathrm{Al_{0.33}Ga_{0.67}As}$ bulk supporting four bound states at $T = 77\,\mathrm{K}$. The value of the Fermi energy was set to $E_\mathrm{F} = -230\,\mathrm{meV}$. The positions of the absorption peaks correspond to the difference of energies between bound levels in the quantum well. Due to the symmetry, only parity changing, odd–even or even–odd, transitions are allowed, with a very small $3 \to 4$ absorption because of the low population of level 3

energy is again set to $E_\mathrm{F} = -230\,\mathrm{meV}$ at $77\,\mathrm{K}$. The calculated bound–bound absorption spectrum is depicted in Fig. 11.9. As already mentioned, due to the symmetry of the structure, only transitions between opposite parity states are allowed, with the position of the main peak corresponding to the difference between the ground and the first excited state energy, $E_2 - E_1 = 45\,\mathrm{meV}$. The satellite peaks at $E_3 - E_2 = 73\,\mathrm{meV}$ and at $E_4 - E_1 = 217\,\mathrm{meV}$ are the result of the $2 \to 3$ and $1 \to 4$ transitions, respectively. Although allowed, absorption from level 3 into level 4 at $E_4 - E_3 = 99\,\mathrm{meV}$ is very weak, because of the fact that level 3 is almost unpopulated at low temperatures.

Consider now the free states (i.e. with E above the potential barrier). The wave functions can be calculated directly for any energy using the expressions in Sections 2.7.1 and 2.7.2, and the in-plane effective mass associated with the even state can now be calculated using equation (11.22), which in the case of a long structure reduces to the mass of the AlGaAs barriers. This is an expected result as the free states are not confined within the thin GaAs well layer. The same conclusion applies for the odd-parity wave function and the corresponding in-plane effective mass.

Finally, the matrix element for bound–free transitions should be calculated. Because of the inversion symmetry, only parity-changing transitions between even bound and odd free states or between odd bound and even free states are allowed and have a non-zero dipole matrix element.

The wave functions of free states are delocalised (see Fig. 11.7) and consequently the dipole matrix elements between the confined bound states and the continuum states are much smaller than for bound–bound transitions. Hence, the values of the bound–free absorptions are clearly way below the bound–bound absorptions in quantum wells where both type of intersubband transitions exists. However, of particular interest are quantum wells that support only one bound state. In such structures, which have practical relevance in infrared photodetectors, intersubband bound–free absorption become an important mechanism.

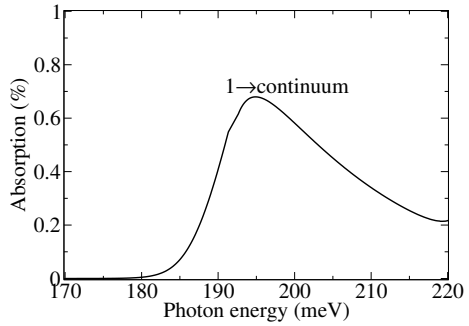


Figure 11.10: Bound–free absorption as a function of the photon energy $\hbar\omega$ for a $4\,\mathrm{nm}$ thick GaAs quantum well in $Al_{0.33}Ga_{0.67}As$ bulk, calculated at 77 K. The quantum well supports only one bound state and all the absorption comes from the bound to continuum transitions. The value of Fermi energy is set to $E_\mathrm{F} = -160\ \mathrm{meV}$

Figure 11.10 displays the calculated bound–free absorption spectra, at 77 K, as a function of the photon energy $\hbar\omega$, again for the rectangular GaAs/$Al_{0.33}Ga_{0.67}As$ narrow quantum well with a 40 Å thick GaAs layer. The structure has only one bound state with an eigenenergy of $E_\mathrm{b} = -183\ \mathrm{meV}$. The doping level is chosen to set the Fermi energy at $E_\mathrm{F} = -160\ \mathrm{meV}$. The absorption line shape yields a step rise at the ionisation threshold $\hbar\omega = |E_\mathrm{b}|$ and a more slowly decaying tail on the high-energy side. The asymmetric bell-like absorption feature arises from the existence of a virtual resonance state in the continuum. This state corresponds to the transmission resonance, when the free state wave function has maximal amplitude in the region above the well. Therefore, if the energy of the photon $\hbar\omega$ approximately coincides with the difference between the energy of the virtual level and the bound state energy, the bound–continuum dipole matrix elements have large values and hence strong absorption peaks.

To summarise, in semiconductor quantum well structures the subband energy levels and the resulting intersubband absorption properties can be specifically tailored with great design freedom. Such structures can, in some sense, be regarded as artificial, human-made atoms or molecules, although they contain two-dimensional subbands and not really discrete energy levels such as quantum dots. But because the joint density of states for intersubband transitions is essentially a δ-function, quantum wells behave like atoms as far as their intersubband absorption properties are concerned. This can be exploited for novel, efficient devices like harmonic generators, or infrared detectors and lasers.

11.7 Intersubband optical nonlinearities

It is well known that quantum-confined structures, like semiconductor quantum wells or superlattices, exhibit a variety of enhanced optical properties compared with bulk materials. An electron in a confining potential is accelerated in this potential under incident laser radiation. The accelerated electron oscillates and radiates according to its motion in the potential. This motion is the origin of the observed optical properties of the material, such

as the refractive index. If the potential is asymmetric and the driving electromagnetic field is large enough to force the electron into the noticeably non-harmonic portion of the potential, then the re-radiated light will contain noticeable fractions of higher harmonics. The actual shape of the effective confining quantum well potential determines the relative strength of the ensuing harmonics, such as in second harmonic generation (SHG) or third harmonic generation. Analogously, an electron in this anharmonic potential driven by two different incident electric fields will lead to sum frequency and/or difference frequency generation of the two incident fields.

Due to the high degree of flexibility in design, with an appropriate choice of materials and structural parameters for optimisation, it is possible to achieve remarkably large values of the second- and third-order nonlinear susceptibilities in semiconductor quantum wells. Optical nonlinearities can be efficiently employed in performing a number of optoelectronic functions—modulation, limiting and thresholding, switching, etc.—maintained over a long interaction length.

Following the seminal work of Gurnick and DeTemple [31] and their prediction that 10–100 times larger second-order optical nonlinearities compared to the lattice value for a specially shaped, graded, asymmetric potential $Al_xGa_{1-x}As$ heterostructure are possible, nonlinear optical effects based on intersubband transitions in semiconductor quantum well structures have been extensively studied, both experimentally and theoretically [32, 33, 34, 35, 36, 37, 38]. These structures possess quite remarkable linear and nonlinear optical properties due to considerable values of the dipole transition matrix elements. A particular nonlinear effect may be grossly enhanced by achieving the resonance conditions [39], when the spacing between relevant bound states coincides with the incident photon energy. Having obtained a resonance in a quantum well, however, the property of interest still depends on the combination of the matrix elements, i.e. via the wave functions, on the potential shape, which may be varied to optimise the structure [40].

Although considerable successes have been demonstrated for intersubband optical nonlinearities, there are several drawbacks worth noting. The nonlinearities are only strongly enhanced for light travelling in the plane of the quantum wells, as significant optical dipole matrix elements only exist for an electric field polarised normal to the quantum well plane, i.e. the plane of electron confinement. As a result, light needs to be efficiently coupled into the in-plane configuration, which is inherently more difficult than normal incidence coupling. Another issue is the resonant absorption accompanying most intersubband optical nonlinearities. The strength of the optical nonlinearity is approximately proportional to the density of electrons in the asymmetric quantum well potential, assuming only one energy level, usually the lowest, is occupied. Energy levels show resonant enhancement of both the nonlinearity and the absorption, with the latter outweighing the former under most conditions. Finally, a challenge is related to phase matching. As semiconductor materials have noticeable non-zero material dispersion across the wide wavelength spans of nonlinear optics, the co-propagation of linear and nonlinear light generally leads to destructive interference of the nonlinearly generated light along the propagation direction of the primary pump beam.

Full analysis of all nonlinear effects is beyond the scope of this chapter; however, the theoretical framework used to calculate the strength of the optical nonlinearities and the expected nonlinear optical output power will be briefly presented for perhaps the most common mechanism—resonant SHG, which enables the generation of coherent radiation

at twice the frequency of available lasers. A full theoretical model employs the density matrix formalism, and the reader can consult, for example, [41] for a detailed derivation. A general theoretical analysis of SHG can then be followed by a calculation of the energy levels and the wave functions from solving Schrödinger's equation in the envelope function approximation for a rectangular asymmetric quantum well structure, in order to optimise effective quantum well potential shapes. The electromagnetic fields are treated classically with Maxwell's equations.

11.8 Electric polarisation

The electromagnetic field of a light wave propagating through a medium exerts a force on loosely bound valence electrons. In normal circumstances this force is weak, and in a linear isotropic medium the resulting polarisation is parallel to and directly proportional to the applied field. In other words, a light beam travelling through the material induces motion of the charged particles that constitute the given medium. In a dielectric medium, the charges are bound together and will start to oscillate in the applied electric field, i.e. they form oscillating electric dipoles. The contributions from the magnetic field part of the light and from electric quadrupoles are much weaker and are usually neglected. This is called the electric dipole approximation. The oscillating dipoles add up to form a macroscopic polarisation P, which is used to describe the material response to the incident field. For low light intensities, i.e. small amplitudes of the electric field E, the charges can follow the field almost exactly and the relationship between P and E is essentially linear, i.e.

$$P = \epsilon_0 \chi E \tag{11.55}$$

where χ is a dimensionless constant known as the dielectric susceptibility. With the invention of lasers, strong fields have become accessible and they are comparable with the fields holding together the atoms in a crystal lattice. For large amplitudes of the field, particle motion will be distorted and equation (11.55) cannot be applied because the nonlinear terms become important. Provided that these new terms are still small compared with the linear one, the polarisation can be expanded as a series in E:

$$P = \epsilon_0(\chi^{(1)}E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \ldots) \tag{11.56}$$

where $\chi^{(1)}$ is the linear susceptibility of the medium while $\chi^{(2)}, \chi^{(3)}, \ldots$ describe the nonlinear optical properties. Given the harmonic electromagnetic field of the input radiation as $E(t) = E_0 \sin(\omega t)$, the quantum well polarisation can be written as:

$$\begin{aligned} P(t) &= \epsilon_0[\chi^{(1)}E_0 \sin(\omega t) + \chi^{(2)}E_0^2 \sin^2(\omega t) + \chi^{(3)}E_0^3 \sin^3(\omega t) + \ldots) \\ &= \epsilon_0\chi^{(1)}E_0 \sin(\omega t) + \frac{1}{2}\epsilon_0\chi^{(2)}E_0^2[1 - \cos(2\omega t)] \\ &\quad + \frac{1}{4}\epsilon_0\chi^{(3)}E_0^3[3\sin(\omega t) - \sin(3\omega t)] + \ldots \end{aligned} \tag{11.57}$$

where the quadratic polarisation term contains a constant static term and a term oscillating at double the incident frequency. The static polarisation produces a dc electric field in the

medium, thus producing an optical rectification. The polarisation oscillating at twice the applied frequency ($2\omega t$) represents SHG characterised with the second-order susceptibility $\chi^{(2)}$. Similarly, the third-order susceptibility $\chi^{(3)}$ characterises third harmonic generation.

11.9 Intersubband second harmonic generation

Consider the general circumstance in which the incident radiation consists of two distinct frequency components ω_0 and ω_1. The strength of the total incident electric field can be represented in the form:

$$E(t) = E_0 \sin(\omega_0 t) + E_1 \sin(\omega_1 t) \tag{11.58}$$

Then assuming that the second-order contribution to the nonlinear polarisation is of the form $P^{(2)}(t) = \epsilon_0 \chi^{(2)} E^2$, the nonlinear polarisation follows as:

$$\begin{aligned} P^{(2)}(t) = \frac{\epsilon_0 \chi^{(2)}}{2} \{ & E_0^2 + E_1^2 - E_0^2 \cos(2\omega_0 t) - E_1^2 \cos(2\omega_1 t) \\ & + 2E_0 E_1 \cos[(\omega_0 - \omega_1)t] + 2E_0 E_1 \cos[(\omega_0 + \omega_1)t] \} \end{aligned} \tag{11.59}$$

The complex amplitudes of the various frequency components of the nonlinear polarisation are therefore given by:

$$\begin{aligned} P(2\omega_0) &= \epsilon_0 \chi^{(2)} E_0^2/2 \\ P(2\omega_1) &= \epsilon_0 \chi^{(2)} E_1^2/2 \\ P(\omega_0 + \omega_1) &= \epsilon_0 \chi^{(2)} E_0 E_1 \\ P(\omega_0 - \omega_1) &= \epsilon_0 \chi^{(2)} E_0 E_1 \\ P(0) &= \epsilon_0 \chi^{(2)} (E_0^2 + E_1^2)/2 \end{aligned} \tag{11.60}$$

In the above expressions $P(2\omega_0)$ and $P(2\omega_1)$ represent SHG. $P(\omega_0 + \omega_1)$ represents sum frequency generation, $P(\omega_0 - \omega_1)$ difference frequency generation and $P(0)$ optical rectification. Out of the four non-zero frequency components, only one will be present in appreciable intensity in typical radiation. The reason for this behaviour is that the nonlinear polarisation can efficiently produce an output signal only if a certain phase matching condition is satisfied, which is usually the case for only one frequency component. In practice, one chooses which frequency component will be radiated by selecting the polarisation of the input radiation and the orientation of the nonlinear crystal.

Now consider an n-doped structure based on direct bandgap semiconductors and take the band gap throughout it to be large enough that interband transitions may be neglected. The polarisation response of the structure to a pump field with photon energy $\hbar\omega$ is then mainly governed by intersubband transitions between quantised conduction-band states. The in-plane dispersion is assumed to be the same for all subbands, i.e. possible differences can be described by the linewidth of the transition. Analogously to the intersubband absorption, intersubband nonlinear susceptibility has a substantial value only for z-polarised radiation.

Taking a realistic n-doped quantum well with its quantised states well separated from each other, one may restrict attention to a few important states (the highly populated ground and a few excited ones). If transitions between these states are approximately resonant with a pair of incident photon energies, $\hbar\omega_0$ and $\hbar\omega_1$, as illustrated in Fig. 11.11, then from the

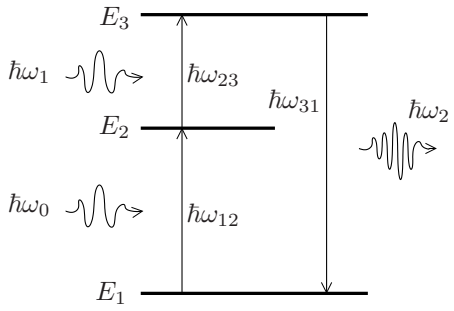


Figure 11.11: Schematic diagram illustrating frequency conversion ($\hbar\omega_2$) in a three-level system from input frequencies $\hbar\omega_0$ and $\hbar\omega_1$

density matrix method [41] applied to a *three-level system*, one finds the expression for $\chi^{(2)}$ at frequency $\omega_2 = \omega_0 + \omega_1$:

$$\chi^{(2)}(\omega_2) = \frac{e^3}{\hbar^2 L_z \epsilon_0}\left(\frac{1}{\omega_{13}+\omega_2-i\gamma_{13}}\right)\left[\frac{(n_2-n_1)d_{12}d_{23}d_{31}}{(\omega_{12}+\omega_0-i\gamma_{12})} - \frac{(n_3-n_2)d_{12}d_{23}d_{31}}{(\omega_{23}+\omega_1-i\gamma_{23})}\right] \tag{11.61}$$

where d_{ij} are the transition dipole matrix elements, L_z is the length of the structure as earlier, n_i are the electron sheet densities (population) in the ground ($i = 1$), first excited ($i = 2$) and second excited state ($i = 3$), ω_{ij} are the transition frequencies (note $\omega_{ij} < 0$ for $i < j$), and γ_{ij} the off-diagonal relaxation rates (i.e. the $i \to j$ transition linewidths). Now, for a monochromatic (laser) pump, i.e. *only one* incident radiation frequency ω exists, set $\omega_0 = \omega_1$ and write it as ω. Also the frequency of interest $\omega_2 = \omega_0 + \omega_1$ becomes 2ω, in which case equation (11.61) transforms into:

$$\chi^{(2)}(\omega_2) = \frac{e^3}{\hbar^2 L_z \epsilon_0}\frac{d_{12}d_{23}d_{31}}{\omega_{13}+2\omega-i\gamma_{13}}\left[\frac{n_2-n_1}{(\omega_{12}+\omega-i\gamma_{12})} - \frac{n_3-n_2}{(\omega_{23}+\omega-i\gamma_{23})}\right] \tag{11.62}$$

Recall that ω_{13}, ω_{12} and ω_{23} are all negative, so the resonance effect as ω (or 2ω in the case of ω_{13}) approaches any of these frequencies is apparent.

In order to maximise nonlinear optical susceptibility and induce efficient SHG in quantum wells, these resonant conditions should be fulfilled, i.e. the difference between subsequent bound state energies should be equal, and at the same time be equal to the incident radiation photon energy, i.e. $E_3 - E_2 = E_2 - E_1 = \hbar\omega$ (also called the *double-resonant condition*). For realistic n-doped quantum wells in thermal equilibrium, only the ground level is highly populated, hence the difference between the electron populations between the second and third excited states, $n_3 - n_2$, in equation (11.62) can be neglected. At exact resonance $\omega_{32} = \omega_{21} = \omega$, which implies that $\omega_{23} = \omega_{12} = -\omega$ and $\omega_{13} = -2\omega$. Taking the off-diagonal relaxation rates to be equal, i.e. $\gamma_{12} = \gamma_{13} \equiv \gamma$, then equation (11.62) simplifies to:

$$\chi^{(2)}_{\text{resonant}} = \frac{e^3(n_1-n_2)}{L_z\epsilon_0}\frac{d_{12}d_{23}d_{31}}{(\hbar\gamma)^2} \tag{11.63}$$

Furthermore, it can also be assumed that population of the first excited state is much lower than population of the ground state ($n_2 \ll n_1$) in which case the total sheet electron density $N_{\rm S}$ (dictated by the doping level) is effectively all in the ground state, i.e. $n_1 = N_{\rm S}$. Hence, equation (11.63) becomes:

$$\chi^{(2)}_{\rm resonant} = \frac{e^3 N_{\rm S}}{L_z \epsilon_0} \frac{d_{12} d_{23} d_{31}}{(\hbar\gamma)^2} \tag{11.64}$$

In conclusion, resonant SHG requires three equidistant energy levels in a quantum well system, with the second-order optical susceptibility proportional to the product of the cyclic dipole matrix elements. In order to further maximise $\chi^{(2)}$, one should clearly maximise the corresponding products of the dipole matrix elements in the numerator of equation (11.64) by appropriate tailoring of the quantum well profile (and hence the wave functions) while preserving the level spacing. The presence of the d_{31} matrix element rules out symmetric quantum wells, as in a parabolic quantum well, because of the definite parity of the wave functions ($d_{31} = 0$), so one should consider asymmetric structures only.

11.10 Maximisation of resonant susceptibility

To find the best potential shape, which will maximise $\chi^{(2)}_{\rm resonant}$, i.e. the product of matrix elements $\Pi^{(2)} = d_{12} d_{23} d_{31}$, the quantum well structure should be varied (i.e. the alloy composition and width of each layer), subject to the constraint that state spacing should be as desired. Various options are possible such as digitally grading the quantum well, using a continually graded quantum well, multiple quantum wells in an external electric field, a step graded quantum well, and a coupled double quantum well. Here the focus will be on the last two cases as they can provide a semi-analytical textbook example of intersubband SHG optimisation.

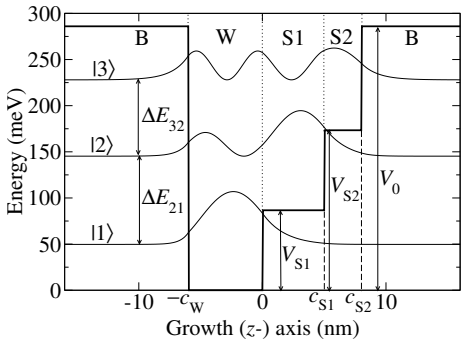


Figure 11.12: The potential (conduction-band edge) in step quantum wells. The structure design parameters, used in the text, are all denoted

Consider an asymmetric step quantum well with a stepwise constant potential and effective mass, as shown in Fig. 11.12. This flexible double-step potential shape can be easily transformed into a simple single-step quantum well by setting $c_{\rm S1} = c_{\rm S2} \equiv c_{\rm S}$ and $V_{\rm S1} =$

$V_{S2} = V_S$, or into an asymmetric double quantum well by choosing $V_{S1} = V_0$, $V_{S2} = 0$, $c_W = c_{W1}$, $c_{S1} = c_B$ and $c_{S2} = c_{W2}$ (see Fig. 11.13), which are frequently used in resonant SHG [36, 39, 40]. As in Chapter 2, the Schrödinger equation should be solved for the quantum well in Fig. 11.12. In the present analysis, depending on the incident photon energies, bound state energies can be some way above the conduction-band minima, and inclusion of the band non-parabolicity, as pointed out in Chapter 3 could be useful.

Having chosen the alloy system to work with (e.g. $Al_xGa_{1-x}As$), it is reasonable to take the well layer to comprise a pure well type semiconductor (GaAs in this instance), because, with dipole matrix elements roughly scaling as $m^{-1/2}$, there is no benefit from allowing the well layer to be made of the alloy. Thus, x_W is defined from the start, and in the step and barrier layers, which are made of the alloy, with suitable compositions x_{S1}, x_{S2} and x_B, the effective mass and potential are uniquely related to each other, i.e. $m_{S1,S2,B} = m_{S1,S2,B}(x_{S1,S2,B})$ and $V_{S1,S2,B} = V_{S1,S2,B}(x_{S1,S2,B})$. Therefore, the energies of the bound states in the system are a nonlinear function of six independent parameters: the widths (coordinates) c_W, c_{S1} and c_{S2} and alloy compositions x_{S1}, x_{S2} and x_B.

More formally, the Hamiltonian for the system is a function of these six variables:

$$\mathcal{H}_0 = \mathcal{H}(c_W, c_{S1}, c_{S2}, x_{S1}, x_{S2}, x_B) \tag{11.65}$$

The system is a viable SHG structure if the double resonance condition is met, i.e. the first three eigenvalues of this Hamiltonian are related by:

$$E_3 - E_2 = E_2 - E_1 = \hbar\omega \tag{11.66}$$

E_1, the ground state energy measured from the well bottom, is left as a free parameter.

The first stage in maximising the resonant susceptibility is, then, to determine a set of viable SHG structures that satisfy the condition above. This can be achieved by specifying the values of four of the six system variables as input parameters, and then solving equation (11.65) iteratively to find the values of the remaining parameters that satisfy the double resonance condition. This solution may be obtained either by an iterative numerical solution of the Schrödinger equation (using the methods of Chapter 3) or by finding a quasi-analytical matching function for the system, using the methods described in Chapter 2. An example of the latter approach is given in [40].

Having found a set of SHG structures (one for each combination of input parameters), the matrix elements can be evaluated for each of the possible structures, either analytically, though via rather cumbersome expressions, or through numerical methods. It is quite straightforward to search the entire free-parameter space and find the best of all quantum wells, which maximises the product of the dipole matrix elements $\Pi^{(2)} = d_{12}d_{23}d_{31}$ in equation (11.64).

Now move from the general quantum well structure shown in Fig. 11.12 to structures based on a single-step quantum well or an asymmetric coupled quantum well as depicted in Fig.11.13. These are technologically more realistic and significantly less demanding to grow than the double-step quantum well. The single-step quantum well in Fig. 11.13(a) can be analysed straightforwardly by modifying the original structure shown in Fig. 11.12 by elimination of the layer S2, i.e. $c_{S1} = c_{S2} \equiv c_S$, and $x_{S1} = x_{S2} \equiv x_S$. The Hamiltonian then simplifies to a function of:

$$\mathcal{H}_0 = \mathcal{H}(c_W, c_S, x_S, x_B) \tag{11.67}$$

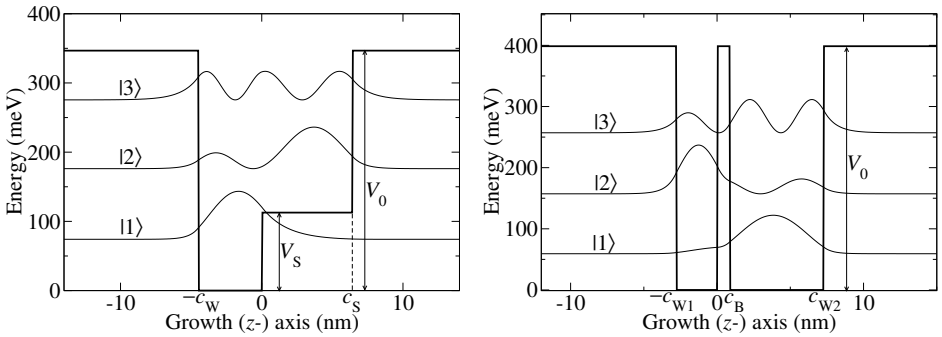


Figure 11.13: The potential (conduction-band edge) of $Al_xGa_{1-x}As$ quantum wells being the targets of optimisation of resonant second harmonic generation, using (a) a single-step geometry and (b) an asymmetric coupled quantum well

Equation (11.67) is greatly simplified and can be solved by specifying only two variables as input parameters. The remaining two parameters and the ground state energy are then found, as before, by solving the equation iteratively such that the eigenvalues satisfy the double resonance condition.

Figure 11.14 illustrates the numerical optimisation of the $Al_xGa_{1-x}As$ single-step quantum well in Fig. 11.13(a). The incident photon energy is chosen to be $\hbar\omega = 100\,\mathrm{meV}$ which, for example, corresponds to a $\lambda \approx 12.4\,\mu\mathrm{m}$ GaAs-based quantum cascade laser operating at room temperature [42, 43]. The width of the well layer c_W and barrier composition x_B were specified as input parameters for each structure. Other parameters (c_S, V_S, E_1) were derived by imposing the double resonance condition on the eigenstates of equation (11.67). The barrier alloy was chosen as the commonly used value $x = 0.4$, resulting in a potential $V_0 = 340\,\mathrm{meV}$.

The largest value of the dipole matrix element product of around $3800\,\text{Å}^3$ was obtained for a well width of $c_\mathrm{W} = 45\,\text{Å}$, as shown in Fig. 11.14(a). Corresponding values of the step width $c_\mathrm{S} = 64\,\text{Å}$ and the step potential of $V_\mathrm{S} = 110\,\mathrm{meV}$ can be found in Fig. 11.14(b).

The results of another example of quantum well optimisation for the same incident photon energy of $\hbar\omega = 100\,\mathrm{meV}$ are presented in Fig. 11.15. This time an asymmetric coupled quantum well structure, shown in Fig. 11.13(b), was analysed. This structure can be obtained by modifying the original structure in Fig. 11.12 using $x_{\mathrm{S}1} = x_\mathrm{B}$, $x_{\mathrm{S}2} = x_\mathrm{W}$, $c_\mathrm{W} = c_{\mathrm{W}1}$, $c_{\mathrm{S}1} = c_\mathrm{B}$ and $c_{\mathrm{S}2} = c_{\mathrm{W}2}$. The Hamiltonian for the system then reduces to:

$$\mathcal{H}_0 = \mathcal{H}(c_{\mathrm{W}1}, c_\mathrm{B}, c_{\mathrm{W}2}, x_\mathrm{B}) \tag{11.68}$$

Again, a unique SHG structure can be found by specifying two of these input parameters, and then imposing the double resonance condition.

Choosing the well width of the first well $c_{\mathrm{W}1}$ as a free parameter and keeping another free parameter, the barrier potential, constant at $V_0 = 400\,\mathrm{meV}$ (aluminium mole fraction in barrier $x = 0.46$) and solving equation (11.68), the maximal product of the matrix elements

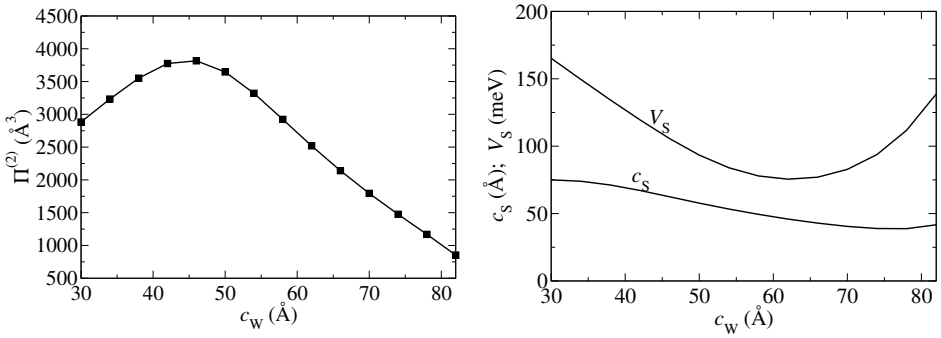


Figure 11.14: Optimisation of an $Al_xGa_{1-x}As$ single-step quantum well under the double resonance condition, $\hbar\omega = 100\,\mathrm{meV}$. The barrier height was kept constant at $V_0 = 340\,\mathrm{meV}$. (a) Matrix element product $\Pi^{(2)}$ and (b) step width and step potential as a function of well width

$\Pi^{(2)} \approx 3900\,\text{Å}^3$ was calculated for $c_{W1} = 28\,\text{Å}$, as seen in Fig. 11.15(a). The optimal thicknesses of other layers of coupled quantum wells, i.e. thin barrier thickness $c_B = 8.5\,\text{Å}$ and the second well width $c_{W2} - c_B = 64\,\text{Å}$, were also obtained automatically during the nonlinear optimisation (see Fig. 11.15(b)).

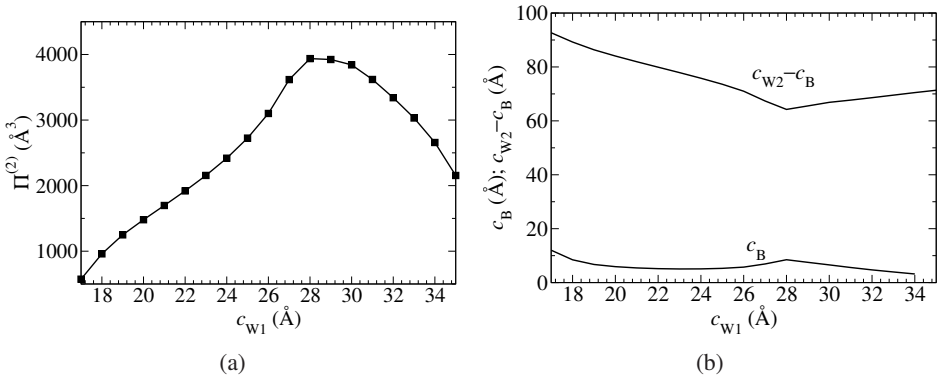


Figure 11.15: Optimisation of $Al_xGa_{1-x}As/GaAs$ asymmetric coupled quantum wells under the double resonance condition, $\hbar\omega = 100\,\mathrm{meV}$. The barrier height was kept constant at $V_0 = 400\,\mathrm{meV}$. (a) Matrix element product $\Pi^{(2)}$ and (b) width of barrier and second well as a function of first well width

The scheme presented for quantum well shape optimisation is quite general. It can be used for other intersubband nonlinear processes, which may not require equispaced states (off-resonant harmonic generation, parametric down-conversion, etc.), hence quantum wells intended for these processes can also be optimised in the same fashion. On the practical side, it may be implemented with reasonable effort and computation time only for structures comprising not more than a few layers of different widths and compositions. Yet it is exactly such simple structures that are of the largest technical importance at present.

Exercises

(1). Starting from equation (11.10), and assuming orthonormality of the wave functions, confirm that the matrix element in equation (11.12) contains only a z-polarised term.

(2). Show that the dipole matrix element $\langle\psi_i|z|\psi_f\rangle$ which appears in equation (11.25) is independent of the choice of origin along the z-axis.

(3). Show analytically, by considering the case of electron states within an infinitely deep quantum well, how the dipole matrix element leads to selection rules for optical intersubband transitions.

(4). Write code to calculate the bound–bound intersubband absorption as given in equation (11.48). By applying it to the case of the single quantum well in Fig. 11.9, show how the application of an electric field breaks the symmetry and gives rise to intersubband transitions that would otherwise not be allowed.

(5). Write code to evaluate the second-order nonlinear susceptibility $\chi^{(2)}$ for the case of second harmonic generation, as a function of photon energy for a single pump source. Again apply it to the case of the single quantum well in Fig. 11.9 under the influence of an electric field. Assume the relaxation rates γ are all equal.

References

[1] R. L. Liboff, *Introductory Quantum Mechanics*, Addison–Wesley, San Francisco, Fourth edition, 2003.

[2] S. Adachi, *GaAs and Related Materials: Bulk Semiconducting and Superlattice Properties*, World Scientific, Singapore, 1994.

[3] P. K. Basu, *Theory of Optical Processes in Semiconductors*, Clarendon Press, Oxford, 1997.

[4] D. D. Coon and R. P. G. Karunasiri, 'New mode of infrared detection using quantum wells', *Appl. Phys. Lett.*, **45**:649, 1984.

[5] P. S. Zory, *Quantum Well Lasers*, Academic Press, Boston, 1993.

[6] W. T. Tsang, Ed., *Lightwave Communications Technology*, Number 22 in Semiconductors and Semimetals. Academic Press, Orlando, 1985.

[7] P. N. J. Dennis, *Photodetectors: An Introduction to Current Technology*, Updates in Applied Physics and Electrical Technology. Plenum Press, New York, 1986.

[8] M. A. Trishenkov, *Detection of Low-Level Optical Signals: Photodetectors, Focal Plane Arrays and Systems*, vol. 4 of *Solid-state Science and Technology Library*, Kluwer Academic Publisher, Boston, 1997.

[9] A. N. Baranov, V. V. Sherstnev, C. Alibert, and A. Krier, 'New III–V semiconductor lasers emitting near 2.6 μm', *J. Appl. Phys.*, **79**(6):3354–3356, 1996.

[10] S. Nakamura and G. Fasol, *The Blue Laser Diode*, Springer, Berlin, 1997.

[11] R. L. Gunshor and A. V. Nurmikko, Eds., *II–VI Blue/Green Light Emitters: Device Physics and Epitaxial Growth*, Number 44 in Semiconductors and semimetals. Academic Press, San Diego, 1997.

[12] E. L. Ivchenko and G. Pikus, *Superlattices and Other Heterostructures: Symmetry and Optical Phenomena*, Springer-Verlag, Berlin, Second edition, 1995.

[13] G. Bastard, *Wave Mechanics Applied to Semiconductor Heterostructures*, Monographies de physique. Halsted Press, New York, 1988.

[14] J. H. Smet, C. G. Fonstad, and Q. Hu, 'Intrawell and interwell intersubband transitions in multiple quantum wells for far-infrared sources', *J. Appl. Phys.*, **79**(12):9305–9320, 1996.

[15] B. K. Ridley, *Quantum Processes in Semiconductors*, Clarendon Press, Oxford, Third edition, 1993.

[16] L. C. West and S. J. Eglash, 'First observation of an extremely large-dipole infrared transition within the conduction band of a GaAs quantum well', *Appl. Phys. Lett.*, **46**:1156, 1985.

[17] D. Kaufman, A. Sa'ar, and N. Kuze, 'Anisotropy, birefringence, and optical-phase retardation related to intersubband transitions in multiple-quantum-well structures', *Appl. Phys. Lett.*, **64**:2543, 1994.

[18] V. Berger, 'Three-level laser based on intersubband transitions in asymmetric quantum wells: a theoretical study', *Semicond. Sci. Technol.*, **9**:1493, 1994.

[19] F. H. Julien, Z. Moussa, P. Boucaud, Y. Lavon, A. Sa'ar, J. Wang, J. P. Leburton, V. Berger, J. Nagle, and R. Planel, 'Intersubband mid-infrared emission in optically pumped quantum wells', *Superlatt. Microstruct.*, **19**:69, 1996.

[20] J. Katz, Y. Zhang, and W. I. Wang, 'Normal incidence infra-red absorption from intersub-band transitions in *p*-type GaInAs/AlInAs quantum wells', *Electr. Lett.*, **28**:932, 1992.

[21] W. Batty and K. A. Shore, 'Normal-incidence TE inter-subband transitions', *IEE Proc. Optoelectron.*, **145**:21, 1998.

[22] C. Sirtori, F. Capasso, J. Faist, and S. Scandolo, 'Nonparabolicity and a sum rule associated with bound-to-bound and bound-to-continuum intersubband transitions in quantum wells', *Phys. Rev. B*, **50**(12):8663–8674, 1994.

[23] S. L. Chuang, *Physics of Optoelectronic Devices*, John Wiley & Sons, Inc., New York, 1995.

[24] P. A. M. Dirac, *The Principles of Quantum Mechanics*, Clarendon Press, Oxford, Fourth edition, 1967.

[25] D. T. F. Marple, 'Refractive index of ZnSe, ZnTe, and CdTe', *J. Appl. Phys.*, **35**:539, 1964.

[26] B. O. Seraphin and H. E. Bennett, 'Optical constants,' in *Optical Properties of III–V Compounds*, R. K. Willardson and A. C. Beer, Eds., number 3 in Semiconductors and Semimetals, pp. 499–543. Academic Press, New York, 1967.

[27] P. Harrison, R. W. Kelsall, P. Kinsler, and K. Donovan, 'Quantum well intersubband transitions as a source of terahertz radiation', in *1998 IEEE Sixth International Conference on Terahertz Electronics Proceedings*, P. Harrison, Ed., 1998, pp. 74–78.

[28] H. Liu and F. Capasso, Eds., *Intersubband Transitions in Quantum Wells: Physics and Device Application 1: Semiconductors and Semimetals*, vol. 62, Academic Press, 2000.

[29] Z. Ikonić, V. Milanović, and D. Tjapkin, 'Bound–free intraband absorption in GaAs–$Al_xGa_{1-x}As$ semiconductor quantum wells', *Appl. Phys. Lett.*, **54**(3):247–249, 1989.

[30] H. C. Liu, 'Dependence of absorption spectrum and responsivity on the upper state position in quantum well intersubband photodetectors', *J. Appl. Phys.*, **73**(6):3062–3067, 1993.

[31] M. K. Gurnick and T. A. DeTemple, 'Synthetic non-linear semiconductors', *IEEE J. Quant. Electr.*, **19**(5):791–794, 1983.

[32] P. Boucaud, F. H. Julien, D. D. Yang, J.-M. Lourtioz, E. Rosencher, P. Bois, and J. Nagle, 'Detailed analysis of second-harmonic generation near 10.6 μm in GaAs/AlGaAs asymmetric quantum wells', *Appl. Phys. Lett.*, **57**(3):215–217, 1990.

[33] C. Sirtori, F. Capasso, D. L. Sivco, A. L. Hutchinson, and A. Y. Cho, 'Resonant Stark tuning of second-order susceptibility in coupled quantum wells', *Appl. Phys. Lett.*, **60**(2):151–153, 1992.

[34] C. Lien, Y. Huang, and T.-F. Lei, 'The double resonant enhancement of optical second harmonic susceptibility in the compositionally asymmetric coupled quantum well', *J. Appl. Phys.*, **75**(4):2177–2183, 1994.

[35] D. Qu, F. Xie, G. Shu, S. Momen, E. Narimanov, C. F. Gmachl, and D. L. Sivco, 'Second-harmonic generation in quantum cascade lasers with electric field and current dependent nonlinear susceptibility', *Appl. Phys. Lett.*, **90**(3):031105, 2007.

[36] E. Rosencher and P. Bois, 'Model system for optical nonlinearities: asymmetric quantum wells', *Phys. Rev. B*, **44**(20):11315–11327, 1991.

[37] D. Indjin, V. Milanović, and Z. Ikonić, 'Application of Bragg-confined semiconductor structures for higher-energy resonant intersubband second-harmonic generation', *Phys. Rev. B*, **55**(15):9722–9730, 1997.

[38] J. Radovanović, G. Todorović, V. Milanović, Z. Ikonić, and D. Indjin, 'Two methods of quantum well profile optimization for maximal nonlinear optical susceptibilities', *Phys. Rev. B*, **63**(11):115327, 2001.

[39] F. Capasso, C. Sirtori, and A. Cho, 'Coupled-quantum-well semiconductors with giant electric-field tunable nonlinear-optical properties in the infrared', *IEEE J. Quant. Electr.*, **30**(5):1313–1326, 1994.

[40] D. Indjin, Z. Ikonić, V. Milanović, and J. Radovanović, 'Optimization of resonant second- and third-order nonlinearities in step and continuously graded semiconductor quantum wells', *IEEE J. Quant. Electr.*, **34**(5):795–802, 1998.

[41] R. W. Boyd, *Nonlinear Optics*, Academic Press, San Diego, 2003.

[42] C. Pflügl, W. Schrenk, S. Anders, G. Strasser, C. Becker, C. Sirtori, Y. Bonetti, and A. Muller, 'High-temperature performance of GaAs-based bound-to-continuum quantum-cascade lasers', *Appl. Phys. Lett.*, **83**(23):4698–4700, 2003.

[43] J. Heinrich, R. Langhans, M. S. Vitiello, G. Scamarcio, D. Indjin, C. A. Evans, Z. Ikonić, P. Harrison, S. Höfling, and A. Forchel, 'Wide wavelength tuning of GaAs/$Al_xGa_{1-x}As$ bound-to-continuum quantum cascade lasers by aluminum content control', *Appl. Phys. Lett.*, **92**(14):141111, 2008.

12

Carrier transport

12.1 Introduction

In Chapter 10 the scattering behaviour of individual carriers was determined for a range of different perturbations, and this enabled the calculation of the *rate* at which carriers transfer between energy subbands in a heterostructure. The focus of this chapter is now to build upon the knowledge of scattering rates, and determine the behaviour of complete heterostructure devices. The first key premise of this discussion is that when an electric field is applied across a structure, scattering rates are enhanced in one direction and the resulting current can be calculated, as described in Section 12.6. Second, in Section 12.5 scattering rates are used to determine the steady-state distribution of carriers, and hence the net population of each subband. In Section 12.9 the population difference between each pair of states is used to predict the net optical gain (or absorption) of radiation within the system.

Clearly, therefore, the scattering behaviour of carriers is key to understanding the electronic and optical behaviour of heterostructure systems. An exemplar heterostructure device, the quantum cascade laser (QCL), is used to illustrate transport phenomena, but the general principles described here can be applied to any system. Finally, Section 12.11 describes how coherent transport (i.e. tunnelling) effects can also be included in a density matrix transport model.

12.2 Quantum cascade lasers

Since their first demonstration in 1994 [1], QCLs have developed rapidly, and are now well established as powerful, compact sources of coherent mid-infrared and terahertz radiation. These devices consist of a periodic set of modules, each containing several quantum wells. When an electric field is applied across the device, electrons 'cascade' between states in each period, emitting a photon at each stage. The QCL is chosen as the focus of study here, since the numerous quantum wells separated by often thin barriers make QCLs an ideal test-bed for studies of quantum mechanical transport in this chapter. Furthermore, this type of laser is *unipolar* in that there is only one carrier type (i.e. electrons), as opposed to the vast majority of laser diodes and light-emitting diodes, which are bipolar and generate light through interband recombination of electrons with holes.

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

The emission wavelength of QCLs is chosen by selecting quantum well widths whose energy spacing corresponds to the desired photon energy, and mid-infrared devices have, to date, been demonstrated throughout the 3–25 μm wavelength band. Very high continuous-wave output powers (several watts) and high wall-plug efficiencies are achievable well above room temperature. QCLs emitting at terahertz frequencies were first demonstrated in 2002 [2], and now cover the 1.2–4.9 THz band. Although high pulsed output powers (greater than 1 W [3]) have been demonstrated, operation of terahertz QCLs is currently limited to cryogenic temperatures [4]. QCL research is a thriving and rapidly developing field, and more detailed discussions of progress in QCLs and their applications can be found in recent review articles (e.g. [5, 6, 7]).

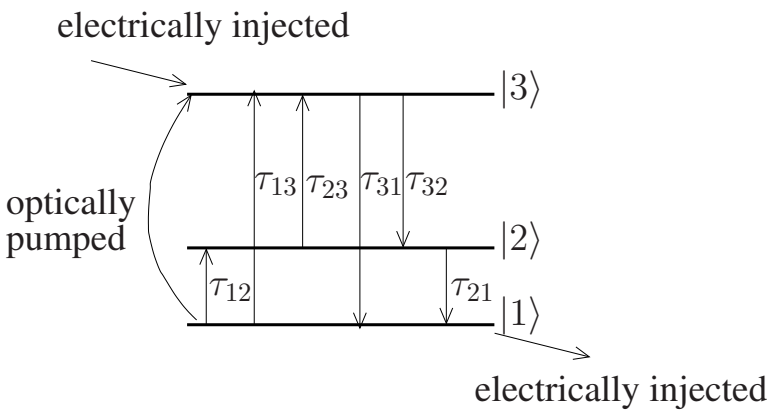


Figure 12.1: The scattering processes in a three-level laser

The simplest active region of any solid state laser would incorporate three energy levels, which in this case would be quantum well subbands, as illustrated schematically in Fig. 12.1. Each level contains a certain number of electrons (say, n_1, n_2 and n_3), which scatter between the subbands (as indicated by the arrows). Energy can be input to the system, either by optically exciting electrons from the ground state $|1\rangle$ to the second excited state $|3\rangle$, or by injecting them directly into the uppermost state. The QCL is an example of the latter, with electrons being removed from the lowest state $|1\rangle$ to be recycled, with these constituting the injected current in an additional stage. Thus, one electron can produce many photons.

Consider the rate equation for the number of electrons (population) of each level, i.e.

$$\frac{\mathrm{d}n_3}{\mathrm{d}t} = I_{\text{in}} + \frac{n_1}{\tau_{13}} + \frac{n_2}{\tau_{23}} - \frac{n_3}{\tau_{32}} - \frac{n_3}{\tau_{31}} \tag{12.1}$$

$$\frac{\mathrm{d}n_2}{\mathrm{d}t} = \frac{n_1}{\tau_{12}} + \frac{n_3}{\tau_{32}} - \frac{n_2}{\tau_{21}} - \frac{n_2}{\tau_{23}} \tag{12.2}$$

$$\frac{\mathrm{d}n_1}{\mathrm{d}t} = \frac{n_3}{\tau_{31}} + \frac{n_2}{\tau_{21}} - \frac{n_1}{\tau_{12}} - \frac{n_1}{\tau_{13}} - I_{\text{out}} \tag{12.3}$$

where I_{in} represents the injection rate (the number of electrons per unit time), which in the steady state is equal to I_{out}. This is a *semi-classical* description as, although the electron energies and wave functions have arisen from solutions of Schrödinger's quantum mechanical wave equation and the scattering rates have been derived from Fermi's golden rule, the interpretation now is one of a flow of particles from one state to another at a calculable rate.

Accepting that this is a model for describing the movement of electrons between quantised energy states, or in this case two-dimensional subbands, implies mathematically that the functions describing the distributions of electron energies within the subbands are all of the same form so that they can be integrated out of the Boltzmann transport equation. It is further assumed that the electron distributions in each subband can be described by a Fermi–Dirac distribution function as in Section 2.5, characterised by some as yet unknown temperature. There are good arguments for this assumption: firstly, calculations in Chapter 10 showed that intrasubband scattering is much, much faster than intersubband scattering and so any events that cause electrons to be scattered into a different subband will quickly cause intrasubband scattering, which will redistribute any energy changes across the subband—driving the subband towards an equilibrium Fermi–Dirac distribution, possibly with a different temperature. This is described as 'thermalisation', and the electron distributions are said to be 'thermalised'. Furthermore, intrasubband scattering mechanisms occur which couple subbands together: electron–electron scattering of the form 21–21[1] are as fast as any other intrasubband events, and so any subband with electrons with a higher average kinetic energy, i.e. a higher temperature, will be driven through these 'bi-intrasubband scattering' events towards thermal equilibrium with the other subbands. So it can also be expected that the subbands will have the same characteristic (electron) temperature [8]; further justification for this is provided at the end of the chapter.

Returning to the rate equations, consider the population of the second level, where in the steady state the net rate of change is zero. Also assume that the temperature is relatively low, and hence the absorption rates can be ignored. Then:

$$\frac{n_3}{\tau_{32}} = \frac{n_2}{\tau_{21}} \tag{12.4}$$

Furthermore, if:

$$\frac{1}{\tau_{21}} > \frac{1}{\tau_{32}} \tag{12.5}$$

then $n_3 > n_2$, i.e. a population inversion will exist between levels $|3\rangle$ and $|2\rangle$, thus fulfilling a necessary condition for stimulated emission. The ratio n_3/n_2 is known as the *population ratio*, which in this analysis would be given by:

$$\frac{n_3}{n_2} = \frac{\tau_{32}}{\tau_{21}} \tag{12.6}$$

Perhaps the simplest way to realise such a three-level system is within a triple quantum well structure, with an energy level in each well—the subband minima can then be altered (almost) independently merely by adjusting the quantum well widths.

[1] This is the shorthand notation of Chapter 10 and means that electrons in initial states 2 and 1 scatter and end up in final states 2 and 1, respectively.

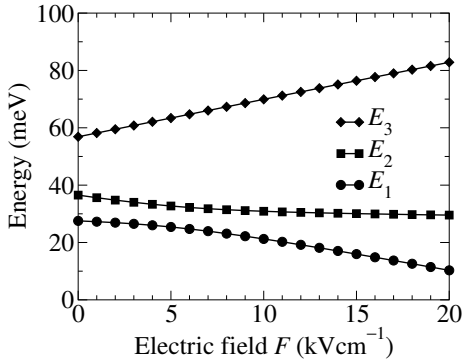


Figure 12.2: The electric field dependence of the lowest three subband minima

Equation (12.5) suggests that considering ways of enhancing the scattering rate from the second level to the first may be a productive way of engineering a population inversion. With this in mind, consider a GaAs triple quantum well surrounded by $Ga_{0.8}Al_{0.2}As$ barriers, with well widths of 56.5, 96.1 and 84.8 Å respectively (integral numbers of monolayers), separated by barriers of width 56.5 and 28.25 Å. The central well has been chosen to be the widest, such that at zero applied electric field it contains the ground state. As the field is increased, an anti-crossing with the state in the right-hand well (84.8 Å) will be inevitable, thus hopefully leading to an increase in the scattering rate, which depopulates level $|2\rangle$.

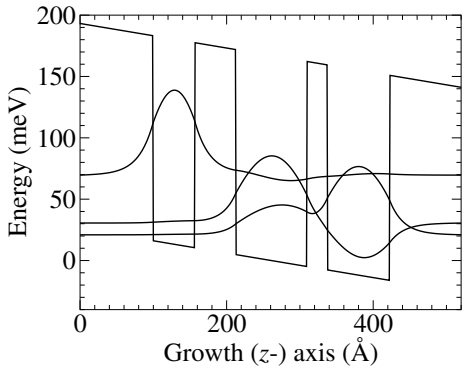


Figure 12.3: A quantum cascade laser active region at $F = 10\,\mathrm{kV\,cm^{-1}}$

This behaviour can be clearly seen in Fig. 12.2, which plots the lowest three subband minima as a function of applied electric field. The wave functions themselves are illustrated in Fig. 12.3, just beyond the anti-crossing, at $F = 10\,\mathrm{kV\,cm^{-1}}$. The large overlap between $|2\rangle$ and $|1\rangle$ is apparent, which will hopefully lead to a strong depopulation of the lower laser ground state, i.e. a high $1/\tau_{21}$. In comparison, the overlap between $|3\rangle$ and $|2\rangle$ is smaller, thus implying a longer carrier lifetime in the upper laser level, i.e. a small $1/\tau_{32}$.

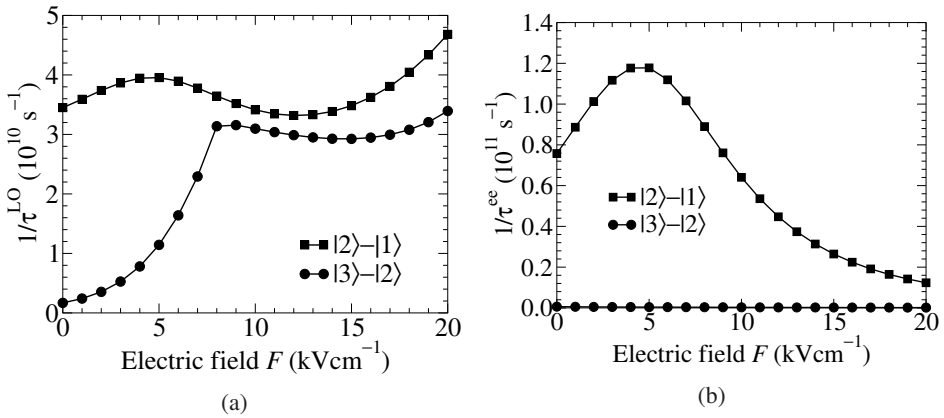


Figure 12.4: The scattering rates from level $|3\rangle$ to $|2\rangle$ and from level $|2\rangle$ to $|1\rangle$, for (a) electron–longitudinal optic phonon processes and (b) electron–electron processes

To confirm whether there is indeed a population inversion requires a calculation of the scattering rates themselves. Using the methods outlined in the previous chapter, Fig. 12.4 displays the electron–LO phonon and the electron–electron scattering rates, as a function of the applied electric field at 77 K and with a carrier density of $10 \times 10^{10}\,\mathrm{cm}^{-2}$ in each level.

The nature of the anti-crossing is evident from the data in Fig. 12.4(a), as the $|3\rangle$ to $|2\rangle$ scattering rate changes rapidly at the anti-crossing, as level $|2\rangle$ moves from being confined in the right-hand well to the central well. This increases the overlap with state $|3\rangle$ confined in the left-hand well, and hence the scattering rate also increases. The depopulation of the lower lasing level, i.e. the rate $|2\rangle$ to $|1\rangle$, mirrors the subband separation between these levels, and indicates that there is mixing between these states for a considerable range of electric fields.

Figure 12.4(b) shows the corresponding electron–electron scattering rates. It is important to note, though, that the main interest here is in finding the net flow of carriers between two specific subbands, rather than the electron–electron scattering rates themselves. For example, a number of different electron–electron processes all contribute to the net flow of carriers from subband $|3\rangle$ to $|2\rangle$. Namely, the 33–23 and 32–22 processes both transfer a single electron between the subbands, while the 33–22 process transfers a pair of electrons. The 32–23 process does not result in a net change of carrier density in the subbands and is, therefore, not counted. There are many additional processes that involve transitions to other subbands: the 33–21, 32–21, 31–23, 31–22 and 31–21 processes all result in a net flow of one electron from $|3\rangle$ to $|2\rangle$. In general, the net rate of carrier transfer from $|i\rangle$ to $|f\rangle$ is given by:

$$\overline{W_{if}} = \overline{W_{iiff}} - \overline{W_{iffi}} + \sum_{j,g} \overline{W_{ijfg}} \tag{12.7}$$

The summation term here includes all the possible initial and final subbands, $|j\rangle$ and $|g\rangle$ respectively, for the second electron involved in the process. The first term 'double-counts' the process that transfers two electrons between the subbands, and the second term cancels out the effect of the process that does not result in a net change of carrier density. Clearly,

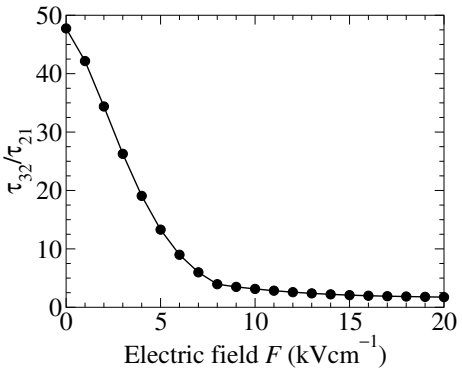


Figure 12.5: The ratio of the lifetimes τ_{32}/τ_{21} of electrons in the triple quantum well active region of the quantum cascade laser

in a laser system with many subbands, the number of processes that must be calculated can become very large, and given that electron–electron scattering rates are rather slow to compute, this can lead to long simulation times. In practice, though, it is often possible to ignore transitions that originate in weakly populated states, or those that involve very large changes in carrier energy.

It is immediately apparent from Fig. 12.4 that, in this case, the electron–electron depopulation rate of level $|2\rangle$, i.e. the rate at which carriers are removed from $|2\rangle$ to $|1\rangle$, is very much faster than the repopulation rate, i.e. the rate at which carriers scatter into $|2\rangle$. This is due to the close proximity in both energy and real space of levels $|1\rangle$ and $|2\rangle$. A more general observation, from comparing Figs. 12.4(a) and (b), shows that the electron–electron scattering rate out of level $|2\rangle$ is very much faster than that due to scattering with LO phonons. In addition, it is apparent that, in both cases, the depopulation rate is larger than the repopulation rate, and hence a population inversion is likely.

This is confirmed by Fig. 12.5, which plots the ratio of the lifetimes—a simple quantity which has been shown to approximate the population ratio between levels $|3\rangle$ and $|2\rangle$. Thus, in conclusion, using the methods outlined, this active-layer design has been shown to exhibit a population inversion between levels $|3\rangle$ and $|2\rangle$ at 77 K. The corresponding emission energy, given by the subband separation $E_3 - E_2$, varies between 33 and $53\,\mathrm{meV}$, which, in turn, correspond to 37 and $23\,\mu\mathrm{m}$, respectively.

It is important to note that these calculations have only estimated the *ratio* between the lifetimes (and hence the populations) in the two excited states so far, and exercise (1) at the end of the chapter explores the implications of these results. However, the optical gain of a laser depends on the *absolute value* of the population difference, and this requires knowledge of the current through the device, which will be considered in the following sections.

A more detailed transport model should account for *all* the scattering processes that were discussed in Chapter 10. Indeed, the extension of wave functions across multiple barriers in a QCL can often result in very rapid interface roughness scattering, which is comparable to or faster than even the carrier–carrier processes. However, as discussed in Chapter 10, interface roughness scattering depends on knowledge of growth-specific parameters (roughness height

and correlation length), and hence, for simplicity, the remainder of this chapter considers only the 'universal' expressions for phonon and carrier–carrier scattering.

12.3 Realistic quantum cascade laser

The above is a simplified model of a QCL to give some insight into the internal workings of those devices. A more realistic model would treat the injection and extraction currents specifically and actually try to calculate them. In order to do this, it would be necessary to model more than one period so that the scattering rates from one period to the next can be calculated. Although most mid-infrared QCLs are based on the InGaAs/InAlAs system, GaAs-based devices have also been demonstrated, and these are more convenient example systems owing to the now familiar set of material parameters. The first GaAs-based mid-infrared QCL, which was developed by Sirtori *et al*. [9], is used as an example in this work. This device has the classic design incorporating a simple active region with three quantum wells providing the ground state, the lower and upper laser levels and a well-defined injector/extractor region consisting of five quantum wells. This is illustrated in Fig. 12.6.

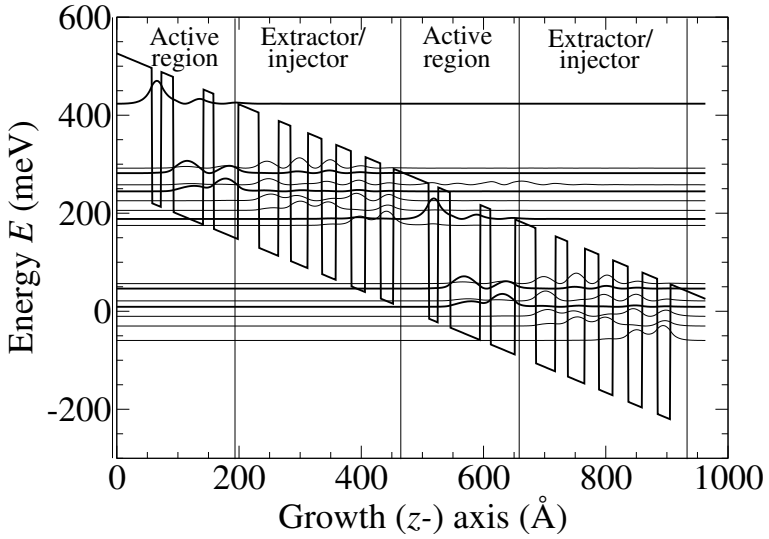


Figure 12.6: A band diagram showing the squares of the wave functions of the electron conduction-band states at a field of $52\,\mathrm{kV\,cm^{-1}}$ for the GaAs/$Ga_{0.67}Al_{0.33}As$ quantum cascade laser of Sirtori *et al*. [9]

Starting at the left-hand edge of the structure, the device has three quantum wells in the active region. The highest energy level is mainly localised in the first quantum well and is the upper laser level, labelled '16' in Fig. 12.7. It is quite separated in both energy and space from the next highest energy level, so the lifetime of carriers in the upper level is long (a few picoseconds!) and a high density of carriers can build up there. The carriers

are eventually scattered out of this level, mainly into the lower laser level, labelled '14' in Fig. 12.7, through both non-radiative (phonon) scattering and stimulated emission within the laser. The third and lowest energy level within the active region, labelled '12' in Fig. 12.7, is usually called the ground state and this is generally designed to have a large spatial overlap with the lower laser level and to have an energy separation of the order of the LO phonon energy ($36\,\mathrm{meV}$ in GaAs) to encourage electrons to scatter out of the lower laser level and into it. As described in Section 12.2, it is this emptying of the lower laser level that produces the population inversion, i.e. more carriers in the higher energy level (16) than in the lower level (14).

In contrast to a traditional diode laser, the electron is still in the conduction band after traversing the active region and generating a photon (or a phonon!), and so the principle of the *cascade* laser is to push this electron into another active region and hence try to encourage it to create another photon, and then another, and so on. This is done from a device design point of view by putting several more states close in energy with reasonable spatial overlap to increase the scattering rate and hence reduce the lifetime of the electrons in the active region ground state. In the example device here, a five quantum well (non-uniform) superlattice is used as illustrated in Fig. 12.6. These 'extractor' states are labelled '9', '10', '11', '13' and '15' in Fig. 12.7. As discussed in Chapter 10, providing multiple states for the electrons to scatter into increases the probability that they will scatter, and that is one of the roles of the extractor. Another role is to provide multiple energy levels close together, certainly separated by less than an LO phonon energy so that the electrons must make diagonally inward transitions, as in Fig. 10.8, which forces them to lose in-plane kinetic energy ($\hbar^2 k_{xy}^2/2m^*$) and settle towards the bottom of the subband: this idea of cooling the carriers is thought to be beneficial for the operation of QCLs, the performance of which does decrease as the temperature increases.

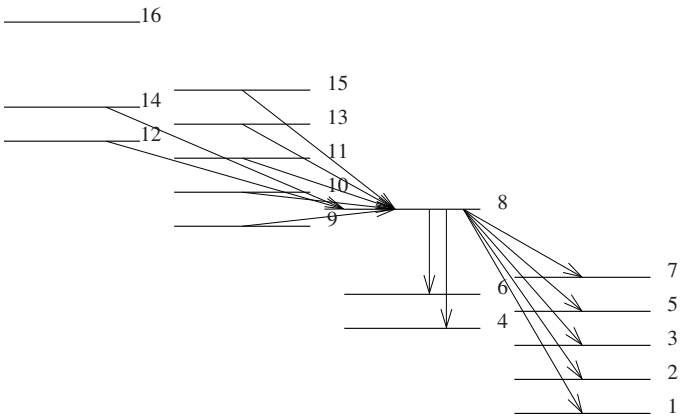


Figure 12.7: A schematic diagram of the energy levels within the two adjacent periods of the quantum cascade laser of Fig. 12.6. Note that the energy levels are numbered over increasing energy within a period. The arrows show all the possible scattering rates that are included in the computation for the example of the upper laser level of the lower-energy (right-hand) period

12.4 Rate equations

The electron densities in each subband of the QCL can again be represented by rate equations as in Section 12.2. However, with many more levels, the rate equations involve more terms. For example, choosing a level near the middle of the structure, i.e. the upper laser level '8' as in Fig. 12.7 and assuming that electrons scatter no further than a complete period, then the rate equation describing the population density (neglecting stimulated emission) is:

$$\begin{aligned}\frac{\mathrm{d}n_8}{\mathrm{d}t} = {} & \frac{n_{15}}{\tau_{15,8}} + \frac{n_{14}}{\tau_{14,8}} + \frac{n_{13}}{\tau_{13,8}} + \frac{n_{12}}{\tau_{12,8}} + \frac{n_{11}}{\tau_{11,8}} + \frac{n_{10}}{\tau_{10,8}} + \frac{n_9}{\tau_{9,8}} + \frac{n_7}{\tau_{7,8}} + \frac{n_6}{\tau_{6,8}} + \frac{n_5}{\tau_{5,8}} \\ & + \frac{n_4}{\tau_{4,8}} + \frac{n_3}{\tau_{3,8}} + \frac{n_2}{\tau_{2,8}} + \frac{n_1}{\tau_{1,8}} - \frac{n_8}{\tau_{8,15}} - \frac{n_8}{\tau_{8,14}} - \frac{n_8}{\tau_{8,13}} - \frac{n_8}{\tau_{8,12}} - \frac{n_8}{\tau_{8,11}} - \frac{n_8}{\tau_{8,10}} \\ & - \frac{n_8}{\tau_{8,9}} - \frac{n_8}{\tau_{8,7}} - \frac{n_8}{\tau_{8,6}} - \frac{n_8}{\tau_{8,5}} - \frac{n_8}{\tau_{8,4}} + \frac{n_8}{\tau_{8,3}} + \frac{n_8}{\tau_{8,2}} + \frac{n_8}{\tau_{8,1}} = 0 \end{aligned} \tag{12.8}$$

which is equated to zero as it is the *steady state* that is of interest. This can be put more succinctly as:

$$\frac{\mathrm{d}n_8}{\mathrm{d}t} = \sum_{j=1}^{15} \left(\frac{n_j}{\tau_{j,8}} - \frac{n_8}{\tau_{8,j}} \right) = 0 \tag{12.9}$$

where in this case there is no need to manually exclude the $j = 8$ term as this will cancel. This is a specific example for state 8 of Fig. 12.7, which now needs generalising to any state in the double period that must be solved. If there are N_{sp} electron energy 'states per period', then equation (12.9) can be written for any state i as:

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} = \sum_{j=i-(N_{\mathrm{sp}}-1)}^{i+(N_{\mathrm{sp}}-1)} \left(\frac{n_j}{\tau_{ji}} - \frac{n_i}{\tau_{ij}} \right) = 0 \tag{12.10}$$

where in this example there are eight states per period (three in the active region and five in the extractor/injector regions) and so, with $N_{\mathrm{sp}} = 8$, the summation over j states goes over seven states of *lower* energy and seven states of *higher* energy than the state in question. In this example the index i varies over the 16 states as indicated in Figs. 12.6 and 12.7. Equation (12.10) thus gives one rate equation for each level in the system, but they are not linearly independent and so it is also necessary to have an additional piece of information in order to derive a unique solution, and the total sheet carrier density per period is introduced and taken as equal to the total sheet doping density:

$$\sum_{i=1}^{N_{\mathrm{sp}}} n_i = \int_{\mathrm{period}} N_{\mathrm{d}}(z)\, \mathrm{d}z \tag{12.11}$$

where $N_{\mathrm{d}}(z)$ is the density of dopant atoms per unit volume, which is usually modulated, hence it is written as a function of position z. Thus the integral of $N_{\mathrm{d}}(z)$ gives the sheet doping density per period. It is usually assumed in QCL modelling that all the donor atoms are ionised.

If the index of the state of interest i is less than N_{sp}, then the summation in equation (12.10) will contain carrier lifetimes τ_{ij} and τ_{ji}, which refer to states that are outside of the two-period domain in Figs. 12.6 and 12.7. In fact, with this approach this happens for all the states in the lower-energy or right-hand period and the converse happens in the higher-energy or left-hand period, as often τ_{ij} and τ_{ji} refer to lifetimes for which i or j is greater than the number of states in the two-period unit cell. To overcome this, periodicity is invoked since:

$$\frac{1}{\tau_{ij}} = \frac{1}{\tau_{i+N_{\text{sp}},j+N_{\text{sp}}}} = \frac{1}{\tau_{i-N_{\text{sp}},j-N_{\text{sp}}}} \tag{12.12}$$

and the indices are shifted by a period to the left or the right until the scattering channel lies completely within the two-period domain. Thus the reason for choosing *two* QCL periods in the model is now clear: it is the *minimum* number that can be used that includes all intra- (within the same) and inter- (between two neighbouring) period scattering channels within the device.

12.5 Self-consistent solution of the rate equations

As discussed in Chapter 10, the lifetimes τ_{ij} and τ_{ji} in equation (12.10) depend on the number of electrons in the initial and final states, i.e. in this case n_i and n_j. So equation (12.10) should be written more correctly as:

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} = \sum_{j=i-(N_{\text{sp}}-1)}^{i+(N_{\text{sp}}-1)} \left(\frac{n_j}{\tau_{ji}(n_j, n_i)} - \frac{n_i}{\tau_{ij}(n_i, n_j)} \right) = 0 \tag{12.13}$$

This can be rearranged to give:

$$\sum_{j=i-(N_{\text{sp}}-1)}^{i+(N_{\text{sp}}-1)} \frac{n_j}{\tau_{ji}(n_j, n_i)} = \sum_{j=i-(N_{\text{sp}}-1)}^{i+(N_{\text{sp}}-1)} \frac{n_i}{\tau_{ij}(n_i, n_j)} \tag{12.14}$$

As n_i is independent of the summation on the right-hand side of the above equation:

$$\sum_{j=i-(N_{\text{sp}}-1)}^{i+(N_{\text{sp}}-1)} \frac{n_j}{\tau_{ji}(n_j, n_i)} = n_i \sum_{j=i-(N_{\text{sp}}-1)}^{i+(N_{\text{sp}}-1)} \frac{1}{\tau_{ij}(n_i, n_j)} \tag{12.15}$$

hence:

$$n_i = \sum_{j=i-(N_{\text{sp}}-1)}^{i+(N_{\text{sp}}-1)} \frac{n_j}{\tau_{ji}(n_j, n_i)} \Bigg/ \sum_{j=i-(N_{\text{sp}}-1)}^{i+(N_{\text{sp}}-1)} \frac{1}{\tau_{ij}(n_i, n_j)} \tag{12.16}$$

Of course, as the lifetimes τ_{ji} and τ_{ij} are complicated functions of the carrier densities n_i and n_j, then it is not possible to completely separate n_i in equation (12.16) and therefore it must be solved self-consistently. One way to do this is to start with initial approximations for the carrier densities n_i and n_j (perhaps take the carrier densities to be equal across all the

subbands) and then use the right-hand side of equation (12.16) to generate a new value for n_i, i.e. interpret equation (12.16) as one of a set of $2N_{\text{sp}}$ iterative equations:

$$n_i^{(n+1)} = \frac{\sum_{j=i-(N_{\text{sp}}-1)}^{i+(N_{\text{sp}}-1)} n_j^{(n)}/(\tau_{ji}(n_j^{(n)}, n_i^{(n)}))}{\sum_{j=i-(N_{\text{sp}}-1)}^{i+(N_{\text{sp}}-1)} 1/(\tau_{ij}(n_i^{(n)}, n_j^{(n)}))} \tag{12.17}$$

where $n_i^{(n)}$ is the nth approximation to n_i. These are solved in sequence for increasing values over the set of subbands i, which then give improved approximations for each of the carrier densities n_i. After each new set of n_i is generated, the necessary additional information of the total carrier density in each period, from equation (12.11), is input by renormalising according to this, i.e.

$$n_i^{(n+1)} \longrightarrow \frac{n_i^{(n+1)}}{\sum_{j=1}^{N_{\text{sp}}} n_j^{(n+1)}} \tag{12.18}$$

This new set of carrier densities is then used to calculate a new set of lifetimes $\tau_{ij}(n_i, n_j)$, which then allows for the calculation of a further improved set of n_i using equation (12.17) and the process is repeated until convergence is achieved, i.e. until the new set of carrier densities $n_i^{(n+1)}$ is equal (within some predefined accuracy) to the previous set $n_i^{(n)}$.

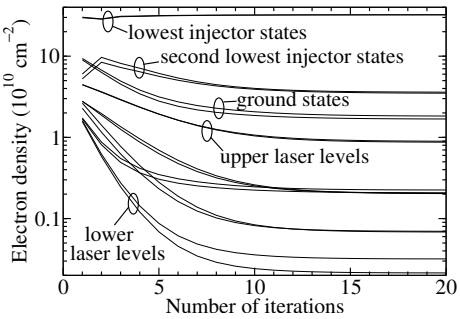


Figure 12.8: The electron densities in each subband through the self-consistent iteration for the mid-infrared quantum cascade laser of interest at a temperature of 77 K and a bias field equivalent to Fig. 12.6 of $52\,\text{kV}\,\text{cm}^{-1}$

Figure 12.8 shows the results of such a self-consistent iteration for the solution of the electron densities in each subband for the QCL in Fig. 12.6. It can be seen from the figure that the electron densities in each level do converge as the number of iterations increases above about 10. Also, the method naturally reproduces the periodicity of the electron populations in equivalent subbands a period apart, which is reassuring and lends weight to the validity of the technique.

The calculations have shown, for this particular device at this temperature and bias, that the subband with the highest population is the lowest state in the extraction/injector region. The second highest populations are to be found in the second lowest injector state, with the third

highest populations being in the lowest-energy state (the so-called ground state) found in the active region (indicated by one of the thicker lines in Fig. 12.6). Note that the total carrier density of $39 \times 10^{10}\,\mathrm{cm}^{-2}$ in each period of the QCL is calculated from the doping density and assumes all the electrons are ionised from the donors. The position of the donors is not included in this calculation as the self-consistent solution of Poisson's equation, as described in Section 3.23, is not included in these calculations. The effect of this on the band-edge potential at this medium carrier density is still quite low (see, for example, [10]).

Figure 12.8 shows quite clearly when self-consistency is reached that there are approximately two orders of magnitude more carriers in the upper laser level than in the lower level laser, which is good because it indicates that there is a strong population inversion as observed in experiment (the device does lase!).

12.6 Calculation of the current density

Once the set of equations (12.17) has been iterated to convergence, then the carrier densities n_i, together with the lifetimes of the carriers in each level in relation to scattering into every other level τ_{ij}, are also known: this is a complete description of the electronic transport within the laser or whatever other device has been studied, and, from this, physical observables such as the current density and gain can be calculated.

The terms like n_i/τ_{ij} represent the number of carriers (in this case electrons) per unit area that make a transition from state i to state j every τ_{ij} seconds. So, it really just represents a current density, i.e. a charge per unit time. A simple model for the current density in the device as a whole is therefore just to take the sum of all such terms which move electrons across some reference plane in the device, as illustrated in Fig. 12.9.

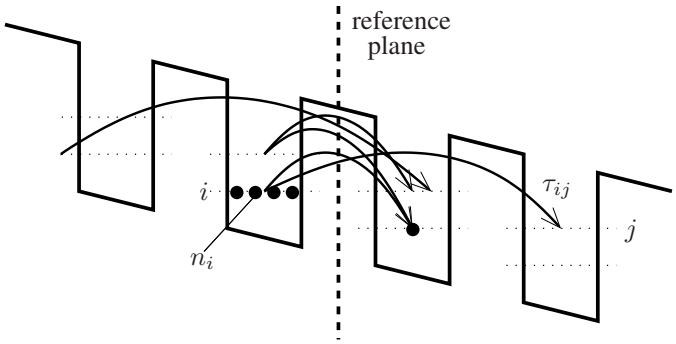


Figure 12.9: Calculating the current density by summing the contributions over all the transitions across a reference plane

For convenience in this case, the reference plane for the current density calculation can be taken as the boundary between the two distinct QCL periods, as illustrated in Figs. 12.6

and 12.7. Hence, in that case the current density would follow as:

$$J = e\left(\sum_{i=N_{\rm sp}+1}^{2N_{\rm sp}}\sum_{j=1}^{N_{\rm sp}}\frac{n_i}{\tau_{ij}} - \frac{n_j}{\tau_{ji}}\right) \tag{12.19}$$

where the second term accounts for back-scattering.

A more detailed model [11] might not assume that the carriers jump the same distance, i.e. a period, in each scattering event and could include the relative separations between the expectation values of the position of the wave functions, but this simple model is fine here.

12.7 Phonon and carrier–carrier scattering transport

Figure 12.10 shows the results of solving for the current density for different bias fields at 77 K for the QCL of interest here, when including LO phonons only, and then adding in acoustic phonons and finally electron–electron scattering. The calculations were performed in this sequence in order to gain an appreciation of their relative contributions to the physical observable of the current density.

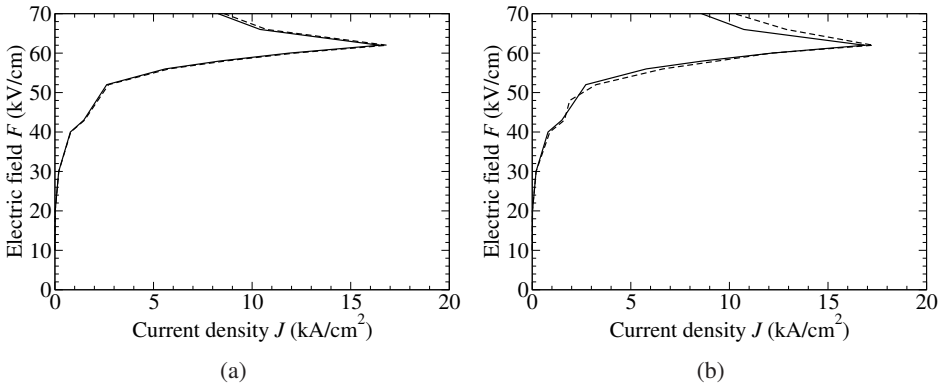


Figure 12.10: The electric field F versus current density J for the quantum cascade laser as above. In (a) the solid line represents the results of calculations including just LO phonon scattering, whereas the dashed line also includes acoustic phonon scattering; in (b) the solid line represents the results of calculations including LO and acoustic phonon scattering, whereas the dashed line also includes electron–electron scattering

Figure 12.10(a) shows the effect of including just phonon scattering mechanisms. For interest, following the procedures discussed above gave 200 different phonon scattering channels, which had to be evaluated for both LO and acoustic phonon scattering. It can be seen that LO phonon scattering dominates the transport in this mid-infrared QCL, with acoustic phonon scattering making an almost negligible contribution. The same can also be said for the 552 electron–electron scattering rates that were evaluated: Fig. 12.10(b) shows

that these also have a small contribution to the current density. These are fairly general conclusions that have been observed in calculations of other devices [12, 13, 14], and the reason is that the intersubband energy separations are typically of the order of a few tens of millielectronvolts and the temperature was mid-range. At very low temperatures, LO phonon scattering at energy spacings less than 10–$20\,\mathrm{meV}$ becomes slow and low-energy scattering such as acoustic and electron–electron will become more important. This is also the case in the longer wavelength terahertz or far-infrared QCLs where the energy gaps and operating temperatures are both usually lower [15, 16, 17].

12.8 Electron temperature

As stated at the beginning of this chapter, this rate equation method assumes that the electron distribution in each subband can be described by a Fermi–Dirac distribution function, and the temperature characterising each distribution is the same. In the calculations that followed it was assumed that this temperature was equal to the lattice temperature. In fact, it is now possible to discard that last assumption and calculate an *average* electron temperature based on an energy balance method. To be more specific, the argument is that, given that any intersubband device will always reach a steady state under continuous operating conditions, the rate at which the electron distributions gain *kinetic* energy (relative to the particular subband minimum) through scattering will balance with the rate at which they lose kinetic energy to the lattice.[2]

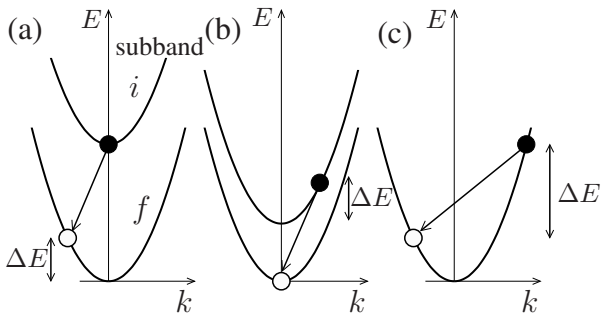


Figure 12.11: Schematic diagram illustrating inelastic (LO phonon) intersubband scattering processes for subband separations (a) greater than and (b) less than the LO phonon energy; (c) illustrates an inelastic intrasubband scattering process

For example, consider the two instances of intersubband scattering as in Fig. 12.11(a) and (b). In Fig. 12.11(a) the subband separation $E_i - E_f$ between the initial electron state $|i\rangle$ and the final state $|f\rangle$ is greater than the dominant phonon energy (in III–Vs the polar optic

[2]The lattice temperature itself will also reach a steady-state value, using electrothermal modelling [18, 19], as discussed in Chapter 13, but for simplicity it will be assumed here that the lattice temperature is constant across the physical dimensions of the device and is equal to the known temperature of some heat-sink.

(LO) phonon), hence any phonon emission event produces a carrier in the lower subband *with more kinetic energy in relation to the subband minimum*. This energy adds to the *total* kinetic energy of the electron distribution but, because of the presence of very rapid sub-picosecond intrasubband carrier–carrier scattering events [8], this additional energy is quickly shared out among the electrons in that subband: the consequence is that the electron distribution within that subband can still be described by a Fermi–Dirac distribution function, albeit with a different temperature. This is a process known as 'rethermalisation'.

Now the electrons in this distribution have more kinetic energy in total than previously and hence the distribution as a whole has a higher temperature than the other subband populations: thermodynamics then drives the system to redistribute this energy. This occurs through 'bi-intrasubband carrier–carrier scattering' (of the form $|j\rangle\,|g\rangle \rightarrow |j\rangle\,|g\rangle$) which redistributes this energy throughout the set of subbands within the quantum device to equalise the temperatures of the distributions [8].

In Fig. 12.11(b) the subband separation $E_i - E_f$ is less than the phonon energy, hence a scattering event from the upper to the lower subband *reduces* the total kinetic energy of the electrons.

In both cases (a) and (b), the change in the total kinetic energy of the electron distributions can be written as $\Delta E = E_i - E_f - E_{\mathrm{LO}}$, where $\Delta E > 0$ should be interpreted as an increase and $\Delta E < 0$ a decrease in this total energy. If there are n_i carriers in the initial state and the LO phonon transitions have associated scattering rates of $1/\tau_{if}^{\mathrm{em.}}$ and $1/\tau_{if}^{\mathrm{abs.}}$ for emission and absorption processes, respectively, then the net kinetic energy generation rate from *intersubband* scattering is:

$$\sum_{f\neq i}\sum_{i}\left[\frac{n_i}{\tau_{if}^{\mathrm{em.}}}(E_i - E_f - E_{\mathrm{LO}}) + \frac{n_i}{\tau_{if}^{\mathrm{abs.}}}(E_i - E_f + E_{\mathrm{LO}})\right]$$

where the indices on the summations imply over all initial and final states in the quantum system. The case when $f = i$ will be discussed below.

One of the main features of QCLs is the bridging (injector) regions between the active layers. These regions serve to remove the electrons from one active region and supply them, at the correct energy, to the next active region. There is, however, another important role that they provide—the opportunity for hot (high kinetic energy) carriers to cool. The injector–collector regions can be optimised to do this by designing the separation of the subbands to be less than the LO phonon energy. Hence, they encourage the carriers to scatter as in Fig. 12.11(b), thus losing kinetic energy and cooling the electron distribution (of course, phonons are generated in this process, which heat the lattice, but as stated above it is assumed here that the device is thermally well connected to a large heat sink, which keeps the lattice temperature constant). Another mechanism that contributes to this cooling is illustrated in Fig. 12.11(c): intrasubband phonon emission. Such transitions lead to a *decrease* in the energy by an amount E_{LO}; again if this has a scattering rate of $1/\tau_{ii}^{\mathrm{em.}}$, then the corresponding kinetic energy loss rate from *intrasubband* scattering is:

$$\sum_{i}\left[\frac{n_i}{\tau_{ii}^{\mathrm{em.}}}E_{\mathrm{LO}} + \frac{n_i}{\tau_{ii}^{\mathrm{abs.}}}(-E_{\mathrm{LO}})\right]$$

where $1/\tau_{ii}^{\text{abs.}}$ accounts for intrasubband phonon reabsorption, which reduces the above energy loss rate by $-E_{\text{LO}}$. This is just the case when $f = i$ as in the previous expression; thus the two will be able to be combined succinctly (see below).

Carrier–carrier (electron–electron or hole–hole) scattering events are described as *elastic*, which means that the total energy of the particles before the event is the same as that after. However, intersubband electron–electron transitions do convert potential energy into kinetic energy (or vice versa), which from the viewpoint of this work would lead to an increase (decrease) in the total kinetic energy of a subband population. Note that the potential energy as defined here includes the quantised component of the kinetic energy, i.e. the energy of the subband minimum.

The expressions for the intersubband and intrasubband cases above for both phonon and carrier–carrier scattering can thus be combined into one expression and, as it has been argued that in the steady state the net kinetic energy generation rate is zero, then:

$$\Delta = \sum_{\text{em.,abs.,c–c}} \sum_{f} \sum_{i} \frac{n_i}{\tau_{if}} (E_i - E_f + \delta E) = 0 \tag{12.20}$$

where the change in energy δE is equal to $-E_{\text{LO}}$ for phonon emission (em.), $+E_{\text{LO}}$ for absorption (abs.) and zero for carrier–carrier (c–c) scattering.

Equation (12.20) is a kinetic energy balance equation: the scattering rates in equation (12.20) are functions of the subband populations n_i and the electron temperatures: if the electron temperature is too low, the number of intrasubband scattering events, as in Fig. 12.11(c), is too small and the kinetic energy equation cannot be balanced. The computation procedure therefore is to vary the electron temperature (which, as argued above, is assumed to be the same for all subbands) until the kinetic energy balance equation is satisfied self-consistently.

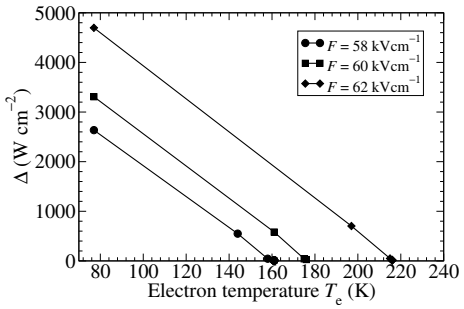


Figure 12.12: The calculated value of the kinetic energy generation rate Δ as a function of the electron temperature and for several different electric fields

Figure 12.12 illustrates this procedure for deducing the electron temperature. Following the standard method for calculating the carrier densities and lifetimes across all levels of the device, the self-consistent iteration is extended to include the effect of the electron temperature. The data in Fig. 12.12 show the variation of the net kinetic energy generation

rate Δ with the electron temperature T_e for the cascade laser of interest here. The solutions for the electron temperatures occur when $\Delta = 0$. Detailed study of the contributions to the electron temperature highlight two interesting points: first, that phonon absorption mechanisms are important even at low (77 K) lattice temperatures' and second, that, for lattice temperatures above 50 K, electron–electron scattering is not an important electron heating mechanism (accounting for less than 1% of the total heat generation rate) in this mid-infrared laser.

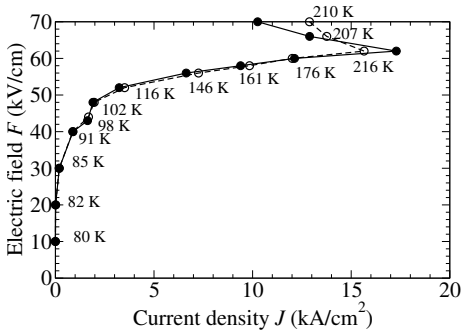


Figure 12.13: The effect of elevated electron temperatures on the electric field F versus current density J for the quantum cascade laser as above. The closed symbols represent the data for the electron temperature fixed at the lattice temperature of 77 K, and the open symbols show the equivalent points where the electron temperature is allowed to vary and is calculated self-consistently. The actual calculated values of T_e are shown for each electric field point

Figure 12.13 shows the results of including the electron temperature in the self-consistent iteration. It can be seen that, although the current density does not change very much from the case when the electron temperature is fixed at the lattice temperature, when allowed to vary, the electron temperature is considerably higher. As might be expected, the electron temperature increases as the current density through the device increases, which is merely a reflection on the increased energy being forced into the device. Figure 12.14 summarises this and shows that indeed there is a linear dependence of the electron temperature on the power density in the device—an effect that has been observed in experiment [20, 21], thus helping to substantiate the assumptions and approach of the model.

12.9 Calculation of the gain

As mentioned in Section 12.6, the self-consistent iteration yields the density of electrons in each quantum well subband and the lifetimes for transitions to every other subband: it is a fundamental description, which allows the physical observable of the current density to be calculated as well as the optical gain of the laser.

Chapter 11 described the absorption of light due to the interaction with carriers within a heterostructure. The strength of the interaction in equation (11.48) depends on the dipole

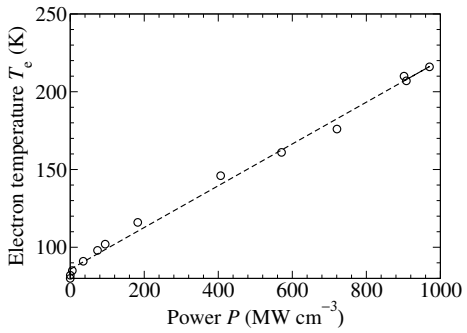


Figure 12.14: The electron temperature T_e plotted as a function of the input power density $F \times J$

matrix element and the difference in the density of carriers in the two subbands $n_i - n_j$. The gain is the exact opposite of absorption, i.e. an *increase* in intensity as light passes through a heterostructure. In fact, it is a manifestation of the same physical light–matter interaction process, except that carriers in excited states transfer their energy into the light field, rather than carriers in the ground state absorbing the energy. As such, equation (11.48), for absorption, can be reused to compute the gain, with the initial state now being higher in energy than the final state:

$$G_{ij}(\omega) = \frac{e^2 \omega \pi}{\overline{n} \epsilon_0 c} |\, \langle \psi_i | z | \psi_j \rangle \,|^2 L_{if}(\hbar\omega, \hbar\omega_0)\,(n_i - n_j) \tag{12.21}$$

The complete gain spectrum can be obtained by summing this expression over every possible pair of subbands in the system. The gain will be positive for any transition in which more carriers exist in the higher-energy state than in the lower-energy state (i.e. a population inversion). Conversely, the gain will be negative (i.e. an absorption) for any transition in which more carriers lie in the lower-energy state. In a QCL, therefore, the complete gain spectrum will exhibit absorption at certain frequencies, and gain at others. In practice, though, the greatest interest lies in calculating the *peak* gain of the system at a specific emission frequency. To a reasonable approximation, the lasing transition in a mid-infrared QCL is quite far in energy from any other pair of states, and it can be considered independently, rather than calculating the gain from a sum over all the transitions in equation (12.21). Also, since a QCL is a periodic system, with photons being generated in every period, it is more useful to describe the gain associated with a single period of the device. A further refinement is obtained by noting that only a fraction of the light actually remains within the laser medium, while the rest escapes into the device substrate. The resulting property, which accounts for all these effects, is the *modal gain* for an optical transition between levels i and j, and is given by [12]:

$$G_\mathrm{M} = \frac{4\pi e^2}{\epsilon_0 \overline{n}} \frac{|\, \langle \psi_i | z | \psi_j \rangle \,|^2}{\gamma_{ij} L_\mathrm{p} \lambda} \Gamma (n_i - n_j) \equiv g \Gamma J \tag{12.22}$$

where λ is the free-space wavelength of the emission, $\overline{n}$ is the mode refractive index, γ_{ij} is the full width at half maximum of the electroluminescence spectrum of the device below

lasing threshold (usually taken from experimental measurements but it can be calculated; see the discussion and references in [11]), L_{p} is the length of one QCL period and Γ is the modal overlap factor (see Chapter 13, particularly equation (13.70)).

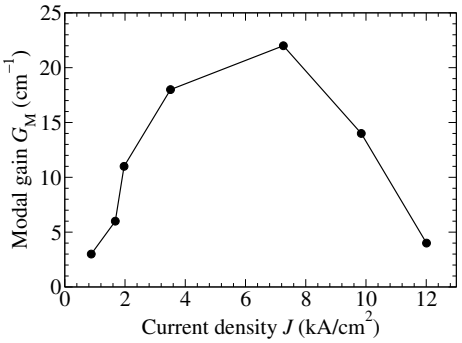


Figure 12.15: The modal gain G_{M} as a function of the current density J

Figure 12.15 illustrates results of calculations of the modal gain as a function of the current density for the QCL of interest in this chapter. It can be seen that the modal gain increases as the amount of current through the device increases, reaches a peak and then decreases. The latter happens even though the current density is still increasing and is due to a decrease in the overlap (the matrix element $\langle\psi_i|z|\psi_j\rangle$) between the upper and lower laser levels as the bias across the device increases. When the current is increased to such an extent that the modal gain is larger than the optical losses of the cavity or waveguide, lasing can occur. This current is known as the 'threshold current' (see Chapter 13 and Section 13.4.2, in particular).

12.10 QCLs, QWIPs, QDIPs and other methods

This rate equation method has been developed further to include self-consistent iteration with Poisson's equation (see Section 3.23) to form a 'self-self-consistent' solution to transport in semiconductor heterostructures. This has been applied successfully to the problem of the effect of doping density on the transport characteristics, namely the current–voltage curves, and to explain the experimentally observed saturation of the maximum current density which occurs for increased doping density [22, 23, 24, 25]. A more sophisticated multi-subband self-consistent model has also been demonstrated, in which the carriers in each subband are still assumed to follow a Fermi–Dirac distribution, but with each subband having an independent carrier temperature [26]. In addition, meaningful predictions have been made relating to the amount of current that escapes from the quantum well system to form a leakage current. This phenomenon, known as 'thermionic emission', has been reported in [27].

The versatility of the rate equation method is also demonstrated by the application to quantum well infrared photodetectors (QWIPs). In the first instance QWIPs seem to have a simpler band structure (see Fig. 12.16); however, the calculations are more complicated as much of the current propagates through the continuum states above quantum well potentials.

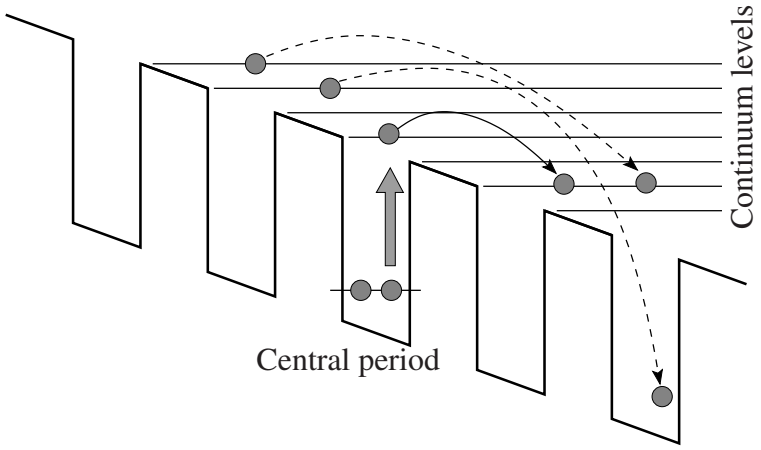


Figure 12.16: Schematic diagram of the electron energy states in a quantum well infrared photodetector. The photo-ionisation process, electron recapture and continuum state currents are illustrated

To calculate the current accurately, the scattering rates must be summed over the continuum states until convergence is achieved. However, this method allows quantitative predictions to be made of the internal (unobservable) physics such as the capture probability and photoconductive gain, as well as physical observables such as the current–voltage curve and the responsivity [28].

Although it needed quite some modification, this general method of rate equations with calculated scattering rates to represent electron densities and the time to scatter between levels has also been applied to quantum dot infrared photodetectors (QDIPs) [29, 30] and its applicability has been confirmed through successful explanation of experimental data [31, 32, 33].

Summarising, this semi-classical method for calculating electron transport properties in semiconductor heterostructures has been applied and validated through comparison with experiment to mid-infrared and far-infrared QCLs, QWIPs and QDIPs. There is another way of effectively solving for the subband population densities *without* making the assumption that the electron distributions in each subband are thermalised, and this is achieved by tracking the momenta of an ensemble of electrons in a stochastic or Monte Carlo simulation. For a simple introduction to stochastic simulations of carrier scattering, see [34]. Simulations of heterostructure devices in similar conditions to those here showed that the carrier distributions are indeed very close to Fermi–Dirac distributions [35, 36] and that the electron temperatures of the subbands are similar [37], both conclusions lending weight to the approach introduced in this chapter.

One important limitation of the model presented so far is that it ignores the effect of the light field upon the carrier density. In reality, whenever a photon is emitted or absorbed, carriers are transferred between subbands, and this in turn affects the gain of the system.

Therefore, the gain of a laser and the population density of each state vary dynamically, and should be solved self-consistently in the time domain, also taking account of the confinement of light within the laser waveguide. Monte Carlo approaches to this problem have been developed, but these can be extremely time-consuming to solve [38]. A much faster (but approximate) method involves solving a reduced and simplified set of rate equations self-consistently with a photon-density calculation [39].

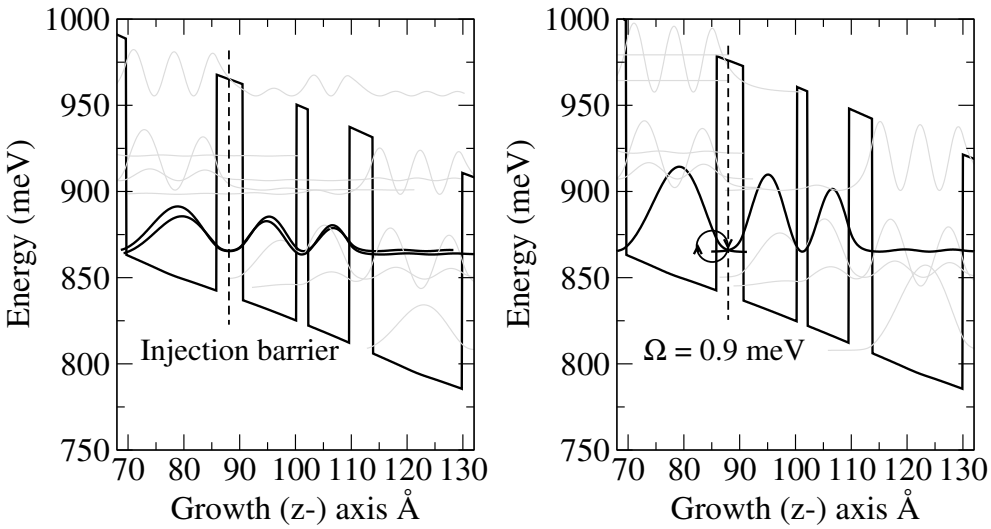


Figure 12.17: Band structure and wave function plot of the three-well terahertz QCL in [40] at $F = 13\,\mathrm{kV\,cm^{-1}}$ with (left) extended wave functions and (right) wave functions calculated with a tight-binding Hamiltonian

12.11 Density matrix approaches

Section contributed by Andrew Grier, University of Leeds

The rate equation method has been shown to give intuitively correct results for charge transport in heterostructure systems such as mid-infrared QCLs. However, it has a fundamental limitation, in that it relies solely on scattering-mediated transport (i.e. a random, incoherent process), and does not consider the *coherent* evolution of the state of the system.

Under certain conditions the rate equation approach may predict physically unrealistic resonances in output parameters since wave functions are allowed to spread over the entire structure, and scattering transport is assumed to be an instantaneous event that occurs after the wave functions have stabilised. Therefore, the rate equation approach effectively allows carriers to be transported instantaneously from one end of the structure to the other. Figure 12.17 (left) illustrates how states calculated with an extended basis can form a symmetric doublet with splitting $\Delta E = 2\hbar\Omega$ when aligned. Since the scattering rate

and therefore rate equation current is dependent on the overlap of states, the current will be independent of barrier thickness. In cases such as terahertz QCLs, injection barriers (the barrier through which electrons are injected to the upper lasing level) are typically thicker than those in mid-infrared QCLs to retain injection selectivity, and current is often overestimated when state alignments occur.

In reality, the injection barriers act as transport 'bottlenecks', and effectively divide the system into finite regions (or modules) in which carriers are 'trapped' for an appreciable time. It is, therefore, more realistic to assume that incoherent scattering only occurs between the quasi-steady states within a single module, and to consider transport *between* the modules as a coherent tunnelling process.

A number of methods exist for modelling coherent quantum transport in semiconductor heterostructures. One of the simplest uses an operator known as a *density matrix* to contain all possible information about a quantum mechanical system in a compact manner. This density matrix approach can deal with two types of uncertainty: the probabilistic interpretation of the wave function described in Chapter 1, and the uncertainty regarding which state the quantum system is in at a given time. The latter can be represented by treating the system as a statistical ensemble of carriers. As in Section 10.2, the state of the system can be represented at any given time as a weighted summation over a number of basis states $|\phi_i\rangle$. These could be any set of orthonormal functions, but it is convenient to choose the states of the unperturbed system. The state of the system $|\psi(t)\rangle$ at any given time is then:

$$|\psi(t)\rangle = \sum_i c_i(t)\,|\phi_i\rangle \tag{12.23}$$

where c_i is the complex amplitude coefficient (or weighting) of the ith state at time t. Now, before defining the density matrix for the system, it is worth studying some of the properties of this sum over basis states.

First, recall that the wave function is normalised (i.e. $\langle\psi(t)|\psi(t)\rangle = 1$) to ensure that the probability of finding a carrier *somewhere* in the system is unity. Substituting this into equation (12.23) gives:

$$\langle\psi(t)|\psi(t)\rangle = \sum_j c_j^*(t)\langle\phi_j| \sum_i c_i(t)\,|\phi_i\rangle \tag{12.24}$$

which can be rearranged to:

$$\langle\psi(t)|\psi(t)\rangle = \sum_j \sum_i c_j^*(t)c_i(t)\langle\phi_j\,|\phi_i\rangle \tag{12.25}$$

This can be simplified by noting that the basis states are orthonormal, $\langle\phi_j|\phi_i\rangle = \delta_{ij}$ (i.e. zero unless $i = j$), and hence one of the summations can be reduced:

$$\langle\psi(t)|\psi(t)\rangle = \sum_j c_j^*(t)c_j(t) = \sum_j |c_j(t)|^2 = 1 \tag{12.26}$$

Therefore, the squared amplitudes of each basis state must add to one at any given time.

Now, consider an observable property, A (position, energy, etc.), of the system, which corresponds to a quantum mechanical operator, $\mathcal{A}$. The expectation value (i.e. average) of the

property is then given by the usual expression:

$$\langle A \rangle = \langle \psi | \mathcal{A} | \psi \rangle \tag{12.27}$$

If the state is resolved into its set of basis states, it is more convenient to consider the operator in its matrix form, with its elements being given by:

$$A_{ij} = \langle \phi_i | \mathcal{A} | \phi_j \rangle \tag{12.28}$$

Substituting equation (12.23) into equation (12.27) gives:

$$\begin{aligned} \langle A(t) \rangle &= \sum_i c_i^*(t) \langle \phi_i | \mathcal{A} \sum_j c_j(t) | \phi_j \rangle \\ &= \sum_i \sum_j c_i^*(t) c_j(t) \langle \phi_i | \mathcal{A} | \phi_j \rangle \\ &= \sum_i \sum_j c_i^*(t) c_j(t) A_{ij} \end{aligned} \tag{12.29}$$

Therefore, it is possible to determine the expectation value of an observable at any given time, provided that the weightings of each state are known at that time.

Recalling now that the weightings can be written as $c_i = \langle \phi_i | \psi \rangle$, the expectation value in equation (12.29) is given by:

$$\langle A(t) \rangle = \sum_i \sum_j \langle \phi_j | \psi(t) \rangle \langle \psi(t) | \phi_i \rangle A_{ij} \tag{12.30}$$

This expression may be simplified by introducing the *density operator*:

$$\rho(t) = |\psi(t)\rangle\langle\psi(t)| \tag{12.31}$$

which describes the probability distribution of the quantum system.

The diagonal elements of the density matrix, ρ_{ii}, represent the probability that a carrier will be found in state i and must satisfy the density conservation law $\mathrm{Tr}(\rho) = 1$. Off-diagonal terms, ρ_{ij}, represent the coherence (or polarisation) between the states:

$$\rho_{ij}(t) = \langle \phi_i | \rho(t) | \phi_j \rangle = c_i(t) c_j^*(t) \tag{12.32}$$

i.e. the coherent superpositions. In a system where carriers can be in several states (such as the subband minima in a QCL), the density matrix will describe the populations of the states as well as the interaction between them. Density matrices have many useful mathematical properties that make them a powerful tool for calculating the properties of systems with mixed states; for a detailed introduction to density matrices and their applications, see [41].

Now, substituting the density operator into equation (12.30) gives:

$$\langle A(t) \rangle = \sum_i \sum_j \langle \phi_j | \rho(t) | \phi_i \rangle A_{ij} = \sum_i \sum_j \rho_{ji}(t) A_{ij} \tag{12.33}$$

This expression may be simplified further by noting that the summation over i gives the product of the elements running along a row of the ρ matrix, and those running down a column of the A matrix. This gives one of the diagonal elements of the matrix $\rho(t)\mathsf{A}$, such that:

$$\langle A(t)\rangle = \sum_j \left[\rho(t)\mathsf{A}\right]_{jj} = \mathrm{Tr}\left[\rho(t)\mathsf{A}\right] \tag{12.34}$$

Therefore, the final result is that the expectation value is the trace (i.e. sum of diagonal elements) of the density matrix multiplied by the matrix operator for the observable.

12.11.1 Time evolution of the density matrix

In practical terms, the key challenge is to determine the unknown values of the density matrix elements at any given time. In typical quantum systems, the statistical distribution of the ensemble of particles will evolve over time, as a response to either an external input or intrinsic oscillations of the system (e.g. Rabi oscillations). In general, the rate of change of the density matrix is given by the derivative of equation (12.31):

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(t) = \frac{\partial}{\partial t}\left[|\psi(t)\rangle\langle\psi(t)|\right] \tag{12.35}$$

Applying the product rule of differentiation to the right-hand side of this expression gives:

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(t) = \left[\frac{\partial}{\partial t}|\psi(t)\rangle\right]\langle\psi(t)| + |\psi(t)\rangle\left[\frac{\partial}{\partial t}\langle\psi(t)|\right] \tag{12.36}$$

Now the terms on the right-hand side can be simplified by noting that the time-dependent Schrödinger equation may be rearranged into the form:

$$\frac{\partial}{\partial t}|\psi(t)\rangle = -\frac{i}{\hbar}\mathcal{H}|\psi(t)\rangle \tag{12.37}$$

and taking the Hermitian transpose of this expression gives:

$$\frac{\partial}{\partial t}\langle\psi(t)| = \frac{i}{\hbar}\langle\psi(t)|\mathcal{H} \tag{12.38}$$

Substituting equations (12.37) and (12.38) into equation (12.36) gives:

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(t) = -\frac{i}{\hbar}\left[\mathcal{H}|\psi(t)\rangle\langle\psi(t)| - |\psi(t)\rangle\langle\psi(t)|\mathcal{H}\right] \tag{12.39}$$

The terms on the right-hand side can be simplified by substituting the density matrix $\rho(t) = |\psi(t)\rangle\langle\psi(t)|$ back into this expression:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\rho(t) &= -\frac{i}{\hbar}\left[\mathsf{H}\rho(t) - \rho(t)\mathsf{H}\right] \\ &= -\frac{i}{\hbar}\left[\mathsf{H}, \rho(t)\right]\end{aligned} \tag{12.40}$$

This is the von Neumann/Liouville equation, which describes the time evolution of the state of a system with a given Hamiltonian. The values of the density matrix elements can, therefore, be found at any given time, and equation (12.34) can then be solved to find the expectation value of any observable property.

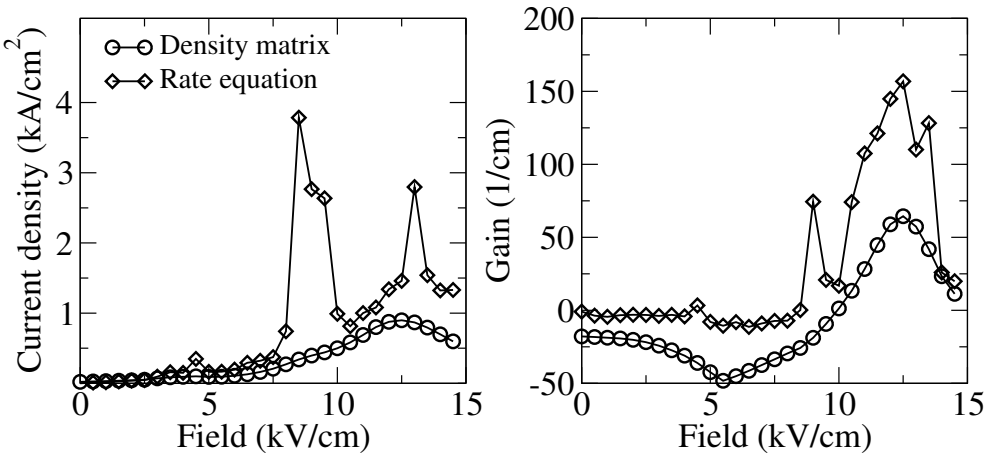


Figure 12.18: Current density (left) and gain (right) calculated with the rate equation and density matrix models for the terahertz QCL in [40] at a lattice temperature of 77 K and electron temperature of 100 K

12.11.2 Density matrix modelling of terahertz quantum cascade lasers

Modelling QCLs with density matrices is a deeper quantum mechanical approach than rate equations since it allows the simultaneous modelling of coherent and incoherent transport. In this approach, it is common to assume that all transport across thick barriers is limited to quantum tunnelling only. Several different density matrix formalisms have used this method to successfully model QCLs such as those in [42, 43, 44].

As discussed previously, rate equation techniques typically consider the eigenstates of an extended potential profile that covers several periods of the QCL structure, as shown in Fig. 12.17 (left). In contrast to this, density matrix approaches typically divide the potential profile of the QCL into individual periods of the structure. The approach of [42] is followed here, which is a general model that does not need *a priori* knowledge of the states present in the system.

The first step of the calculation is to solve the Schrödinger equation separately for three adjacent periods of the QCL, with potential profiles labelled $V_{\mathrm{U}}(z)$, $V_{\mathrm{C}}(z)$ and $V_{\mathrm{D}}(z)$ ('upstream', 'central' and 'downstream', respectively). The edges of each potential profile are 'padded' with thick layers of the barrier material (with potential offset V_0 at the interface) such that the wave functions at the edges of each period decay into the barriers rather than spreading across multiple periods, as shown in Fig. 12.17 (right). The reason for this 'tight-binding' approach is to establish a basis for tunnelling transport between periods of the system, such that the carriers initially located in the upstream states are transported to the central states and from central to downstream states.

The first step in solving the Liouville equation (12.40) for the system is to find the elements of the Hamiltonian matrix. Since the eigenstates were found for each period (using different Hamiltonians), the solutions of the Schrödinger equation for each period are denoted using

a distinct subscript. For example, the wave functions for the central period may be denoted $\phi_{\mathrm{C},1}(z), \ldots, \phi_{\mathrm{C},N}(z)$ with corresponding energies $E_{\mathrm{C},1}, \ldots, E_{\mathrm{C},N}$, while the states in the upstream period are given the subscript U.

If each period of the QCL contains N bound states, the complete Hamiltonian matrix to be used in the Liouville equation will contain $3N \times 3N$ elements, with each describing the interaction between any pair of states in the three-period system. Rather than writing the Hamiltonian matrix in full, it can be written in the block form:

$$\mathsf{H} = \begin{pmatrix} \mathsf{H}_{\mathrm{UU}} & \mathsf{H}_{\mathrm{UC}} & \mathsf{H}_{\mathrm{UD}} \\ \mathsf{H}_{\mathrm{CU}} & \mathsf{H}_{\mathrm{CC}} & \mathsf{H}_{\mathrm{CD}} \\ \mathsf{H}_{\mathrm{DU}} & \mathsf{H}_{\mathrm{DC}} & \mathsf{H}_{\mathrm{DD}} \end{pmatrix} \tag{12.41}$$

where each 'element' is itself an $N \times N$ matrix containing all the interactions between states in the periods denoted by the subscripts. For example, the H_{UC} block describes all the interactions between states in the upstream period and those in the central period, with its elements being given by:

$$H_{\mathrm{UC},ij} = \int \phi^*_{\mathrm{U},i}(z)\ \mathcal{H}\ \phi_{\mathrm{C},j}(z)\ \mathrm{d}z \tag{12.42}$$

i.e. the overlap integral between the two states and the Hamiltonian operator for the system.

A degree of care is required, though, since the Hamiltonian operator $\mathcal{H}$ here refers to the complete three-period system, whereas the basis states in each block of this tight-binding model were found for *different* Hamiltonians using spatially separated potentials. This introduces two complications to the analysis: first, the basis states are eigenvalues of the separated periods (not $\mathcal{H}$); and second, basis states in different periods are not orthogonal to each other. As such, a number of approximations are generally introduced when formulating the Hamiltonian matrix for the system. In the case of the elements within the intraperiod blocks (i.e. UU, CC and DD), the states *are* orthonormal to each other, and are quite close to the eigenvalues of the complete system. Therefore the approximation is made (e.g. in the CC block) that $\mathcal{H}\phi_{\mathrm{C},j}(z) \approx E_{\mathrm{C},j}\ \phi_{\mathrm{C},j}(z)$, and therefore:

$$H_{\mathrm{CC},ij} \approx E_{\mathrm{C},j} \int \phi^*_{\mathrm{C},i}(z)\ \phi_{\mathrm{C},j}(z)\ \mathrm{d}z = E_{\mathrm{C},j}\delta_{ij} \tag{12.43}$$

i.e. the diagonal terms of the H_{CC} matrix block are simply the energies of each state (i.e. $H_{\mathrm{CC},11} = E_{\mathrm{C},1}, H_{\mathrm{CC},22} = E_{\mathrm{C},2}$, etc.), and every other term is zero. Similarly, the energies of the upstream and downstream states are placed on the diagonal elements of the H_{UU} and H_{DD} blocks, respectively.

This approximation cannot be made for the interperiod blocks (UC, CD etc.) since the basis states are not orthonormal and the integral does not reduce to a delta function. However, a coupling strength can be derived for any given transition, i.e. the effective potential that describes the degree to which states in different periods interact across a barrier. This is derived in full in [45], and can be represented by Hamiltonian matrix elements of the form (e.g. for central–upstream transitions):

$$H_{\mathrm{CU},ij} \approx \sqrt{\langle\phi_{\mathrm{C},i}|V_{\mathrm{C}}(z) - V_{\mathrm{ext}}(z)|\phi_{\mathrm{U},j}\rangle \times \langle\phi_{\mathrm{C},i}|V_{\mathrm{U}}(z) - V_{\mathrm{ext}}(z)|\phi_{\mathrm{U},j}\rangle} \tag{12.44}$$

where $V_{\mathrm{C}}(z)$ and $V_{\mathrm{U}}(z)$ are the 'padded' potential profiles for the central and upstream periods of the structure respectively, and $V_{\mathrm{ext}}(z)$ is the potential profile for the complete, extended three-period system.

It is worth noting from equation (12.44) that the coupling strength between states with large spatial separation is very low, and as such it can be assumed that no direct interaction exists between upstream and downstream states. Therefore, every element of the H_{UD} and H_{DU} blocks can be set to zero. Also, since the system is translationally invariant, the interaction between any pair of adjacent periods is identical, so $\mathsf{H}_{\mathrm{UC}} = \mathsf{H}_{\mathrm{CD}}$ and $\mathsf{H}_{\mathrm{CU}} = \mathsf{H}_{\mathrm{DC}}$. However, the diagonal (UU, DD, CC) blocks of the Hamiltonian matrix are different from each other, since the applied electric field shifts the states in each period downward in energy. The Hamiltonian matrix can, therefore, be simplified considerably to:

$$\mathsf{H} = \begin{pmatrix} \mathsf{H}_{\mathrm{UU}} & \mathsf{H}_{\mathrm{UC}} & 0 \\ \mathsf{H}_{\mathrm{CU}} & \mathsf{H}_{\mathrm{CC}} & \mathsf{H}_{\mathrm{UC}} \\ 0 & \mathsf{H}_{\mathrm{CU}} & \mathsf{H}_{\mathrm{DD}} \end{pmatrix} \tag{12.45}$$

All the Hamiltonian matrix elements for the unperturbed system have now been defined, with the intra-period terms giving the state energies and the inter-period terms giving the coupling strengths. This is sufficient to calculate the unperturbed electronic behaviour of the system, but additional terms are needed to describe the interaction of carriers with light. In the simplest models, a pair of states (within the same period) is considered to interact with a light field of angular frequency ω only if $|E_{\mathrm{C},i} - E_{\mathrm{C},j}| \approx \hbar\omega$ (and similarly for the upstream and downstream periods). The off-diagonal intra-period Hamiltonian matrix elements for these optically-active transitions can then be set to:

$$H_{\mathrm{CC},ij} = z_{\mathrm{CC},ij} e A_0 \exp(i\omega t) \tag{12.46}$$

where $z_{\mathrm{CC},ij}$ is the dipole matrix element for the transition, and A_0 is the electric field strength of the incident light field.

The Hamiltonian matrix has now been defined completely, and the Liouville equation may be solved. The density matrix (to be determined) can be written in the block form:

$$\rho = \begin{pmatrix} \rho_{\mathrm{UU}} & \rho_{\mathrm{UC}} & \rho_{\mathrm{UD}} \\ \rho_{\mathrm{CU}} & \rho_{\mathrm{CC}} & \rho_{\mathrm{CD}} \\ \rho_{\mathrm{DU}} & \rho_{\mathrm{DC}} & \rho_{\mathrm{DD}} \end{pmatrix} \tag{12.47}$$

where each block is (as with the Hamiltonian matrix) an $N \times N$ matrix, which represents the coherences between all pairs of states within the same period (UU, CC and DD) or in two different periods (CU, UC, etc.).

As before, the calculation can be simplified by considering the translational invariance of states in a QCL. As such, the coherences between equivalent pairs of states within any period are identical and therefore $\rho_{\mathrm{CC}} = \rho_{\mathrm{UU}} = \rho_{\mathrm{DD}}$. The same applies for equivalent interactions between any two periods, so $\rho_{\mathrm{CD}} = \rho_{\mathrm{UC}}$ and $\rho_{\mathrm{DC}} = \rho_{\mathrm{CU}}$. Complexity is reduced further by again noting that direct interactions between upstream and downstream states are extremely weak. Therefore, $\rho_{\mathrm{UD}} = \rho_{\mathrm{DU}} = 0$ and the simplified density matrix then takes the form:

$$\rho = \begin{pmatrix} \rho_{\mathrm{CC}} & \rho_{\mathrm{UC}} & 0 \\ \rho_{\mathrm{CU}} & \rho_{\mathrm{CC}} & \rho_{\mathrm{UC}} \\ 0 & \rho_{\mathrm{CU}} & \rho_{\mathrm{CC}} \end{pmatrix} \tag{12.48}$$

In the equation of motion given by equation (12.40), scattering and dephasing caused by interaction with ionised impurities, LO phonons, acoustic phonons, alloy disorder and electron–electron scattering will add extra relaxation terms to the diagonal and off-diagonal elements of the density matrix. Equation (12.40) then becomes:

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = -\frac{i}{\hbar}\left[\mathsf{H}(t), \rho(t)\right] - \frac{\rho}{\tau} \tag{12.49}$$

where the final term is the *relaxation matrix* that describes the damping of the system. The diagonal terms of this matrix describe the change in population over time caused by scattering between states, and are given by:

$$\left(\frac{\rho}{\tau}\right)_{ii} = -\frac{\rho_{ii}}{\tau_i} + \sum_{j \neq i} \frac{\rho_{jj}}{\tau_{ji}} \tag{12.50}$$

i.e. the difference between the density of carriers scattering into the state and those scattering out. The off-diagonal elements reflect that every scattering event involving either state i or j will affect the coherence of that state with every other state, and are given by:

$$\left(\frac{\rho}{\tau}\right)_{ij} = \frac{\rho_{ij}}{\tau_{\parallel,ij}} \tag{12.51}$$

where the dephasing time $\tau_{\parallel,ij}$ is calculated as:

$$\frac{1}{\tau_{\parallel,ij}} = \frac{1}{2\tau_i} + \frac{1}{2\tau_j} + \frac{1}{\tau_{ii}} + \frac{1}{\tau_{jj}} \tag{12.52}$$

The reduced contribution from intersubband events is caused by reduction of the decay rate with increasing energy density ($\rho_{ii} \propto \mathrm{e}^{-t/T_1} \Rightarrow \rho_{ij} \propto \mathrm{e}^{-t/2T_1}$). For further discussion of calculating $\tau_{\parallel,ij}$, see [46, 47].

Having now defined all the terms in the Liouville equation (12.49), the next challenge is to solve the equation to find all the density terms at a given time. One approach is to define an initial density matrix for the system and then propagate forward in time using standard time-domain approaches to partial differential equations, e.g. Runge–Kutta methods. However, this can be time-consuming and it is often only the final steady-state condition of the QCL that is of interest. Under this simpler approach, it can be assumed that every density matrix element eventually settles to a form containing only a constant dc component ρ_{ij}^0 and an oscillating component that responds to the light field with frequency ω such that:

$$\rho_{ij}(t) = \rho_{ij}^0 + \rho_{ij}^+ \exp(i\omega t) + \rho_{ij}^- \exp(-i\omega t) \tag{12.53}$$

where $\rho_{ij}^{\pm}$ are constant amplitudes for the forward and reverse rotating components. Strictly, higher-frequency harmonic components will also appear within the density matrix but for simplicity this can be assumed to be a weak effect and the single frequency approximation may be used. The derivative of each density matrix element at steady state is then given by:

$$\frac{\partial \rho_{ij}}{\partial t}(t) = i\omega\rho_{ij}^+ \exp(i\omega t) - i\omega\rho_{ij}^- \exp(-i\omega t) \tag{12.54}$$

A further simplification can be obtained by assuming that the diagonal terms (i.e. the state populations) will be dominated by their dc components. Equation (12.49) may now be multiplied out to obtain a set of linear equations containing the dc and ac coefficients for each density matrix element which can be solved using standard linear algebra software libraries.

Physical quantities such as current density for this device can be extracted from the solved density matrix as $j = \mathrm{Tr}(\rho J)$, with the current matrix derived from the average drift velocity:

$$J = e\frac{i}{\hbar}\left[\mathsf{H}, \mathsf{z}\right] \tag{12.55}$$

The gain of the QCL can also be extracted from J by considering the harmonic (ac) response of density matrix terms to a light field. The complex permittivity $\tilde{\epsilon}$ of the electron gas is calculated from:

$$\tilde{\epsilon} = \frac{j_{\mathrm{ac}}}{\frac{\mathrm{d}}{\mathrm{d}t}\left(A_0 \exp i\omega t\right)} \tag{12.56}$$

and the gain follows as $g = \omega\Im(\tilde{\epsilon})/n_{\mathrm{r}}c$, where n_{r} is the refractive index of the structure. For a more complete explanation of this approach, see [42, 43, 44].

Figure 12.18 shows a comparison of the rate equation and density matrix models applied to the three-well terahertz quantum cascade laser in [40]. The effect of including dephasing and coherent transport across the injection barrier is to reduce the overall current density and suppress the unphysical resonances illustrated in Fig. 12.17. Other examples of applications of density matrix models applied to intersubband devices are QCLs in a magnetic field [48], and quantum cascade detectors [49]. Beyond QCL density matrix models, such as that presented here, lie non-equilibrium Green's functions (NEGFs). In density matrix and rate equation models lineshapes are imposed after calculating densities/populations, but these are an intrinsic part of the Hamiltonian. Since Hamiltonian perturbations cause energy shifts as well as scattering, transitions are broadened and this is accounted for intrinsically in NEGF models, giving an energy-dependent density function $\rho(E)$. Details of NEGF QCL models can be found in [50, 51].

Exercises

(1). Figure 12.5 shows that the *ratio* of carrier lifetime in the simulated three-level QCL increases as the electric field decreases towards zero.

- (a) Explain qualitatively what this implies about the relative population in each state, and whether this is a physically realistic result.
- (b) Use the rate equations for the three-level system to determine the absolute values of the populations when the current is zero. Assume that the absorption rates are negligible. Discuss the significance of this result.
- (c) The optical gain of the system can be determined from the population difference between states $|3\rangle$ and $|2\rangle$ using equation (12.22). Use the rate equations to determine expressions for the gain (i) as a function of current in an electrically pumped system, and (ii) as a function of the pump rate $1/\tau_{13} = W_{\mathrm{p}}$ in an optically pumped system.

(2). Solve the rate equations for an electrically pumped three-level QCL system to find the populations in each of the states as a function of the scattering injection rate W_{I} between periods of the structure. The current density is related to the scattering injection rate by $J = en_1 W_{\mathrm{I}}$. Let the total population be $10 \times 10^{10}\,\mathrm{cm}^{-2}$, the scattering rates for the system be $W_{32} = W_{31} = 1 \times 10^{10}\,\mathrm{s}^{-1}$ and $W_{21} = 10 \times 10^{10}\,\mathrm{s}^{-1}$, and assume that absorption processes are negligible. Determine the populations of the states as the injection rate tends towards zero and towards infinity.

(3). Write a simulation script, using the QWWAD simulation tools, to determine the interface roughness and alloy disorder scattering rates as a function of bias in the three-level system of Fig. 12.3. Compare the simulated rates with those presented in Fig. 12.4 and comment on their significance.

(4). Explain how the populations and temperatures of the initial and final subbands affect the scattering lifetimes. From this basis, comment qualitatively on how the number of iterations needed for self-consistent solutions of the rate equations varies with the doping density in a system.

(5). Assuming that it takes $5\,\mathrm{s}$ to compute the scattering rate for each electron–electron process to high precision, estimate the computation runtime that would be needed to determine the populations self-consistently in the eight-level QCL of Fig. 12.6. Suggest ways in which this could be reduced; estimate the time impact of the proposed computation optimisation schemes and discuss their relative advantages and disadvantages. Note that the number of combinations of pairs of states selected from n different bands is obtained from set theory as:

$$n_{\mathrm{pairs}} = \frac{(n+1)!}{2(n-1)!} \tag{12.57}$$

(6). Given that the peak value of the Lorentzian function is $L(\hbar\omega_0, \hbar\omega_0) = 2/(\pi\gamma_{ij})$, derive the modal gain (equation (12.22)) in a period of a quantum cascade laser, starting from the expression for intersubband absorption (equation (11.48)).

(7). Derive an approximate expression for the gain coefficient g in a three-level electrically pumped laser in terms of the scattering lifetimes. The simplest laser models usually consider only one transition as contributing to the gain, and the gain coefficient is usually treated as a constant parameter. Discuss the validity of these approximations.

(8). Given a Hamiltonian matrix and a density matrix of the form:

$$\mathsf{H} = \begin{pmatrix} E_1 & 0 & 0 \\ 0 & E_2 & 0 \\ 0 & 0 & E_3 \end{pmatrix} \tag{12.58}$$

$$\rho = \begin{pmatrix} \rho_{11} & \rho_{12} & \rho_{13} \\ \rho_{21} & \rho_{22} & \rho_{23} \\ \rho_{31} & \rho_{32} & \rho_{33} \end{pmatrix} \tag{12.59}$$

using the Liouville equation and equation (12.50), derive the three coupled equations necessary to find the diagonal density matrix elements. What do the diagonal terms of

the density matrix represent and what condition must be met to conserve probability density? Which terms of the density matrix above will determine the gain linewidth in QCL simulations?

References

[1] J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho, 'Quantum cascade laser', *Science*, **264**(5158):553–556, 1994.

[2] R. Köhler, A. Tredicucci, F. Beltram, H. E. Beere, E. H. Linfield, A. G. Davies, D. A. Ritchie, R. C. Iotti, and F. Rossi, 'Terahertz semiconductor-heterostructure laser', *Nature*, **417**:156–159, 2002.

[3] L. Li, L. Chen, J. Zhu, J. Freeman, P. Dean, A. Valavanis, A. G. Davies, and E. H. Linfield, 'Terahertz quantum cascade lasers with >1 W output powers', *Electron. Lett.*, **50**(4):309–311, 2014.

[4] S. Fathololoumi, E. Dupont, C. Chan, Z. Wasilewski, S. Laframboise, D. Ban, A. Matyas, C. Jirauschek, Q. Hu, and H. C. Liu, 'Terahertz quantum cascade lasers operating up to ˜200 K with optimized oscillator strength and improved injection tunneling', *Opt. Express*, **20**(4):3866–3876, 2012.

[5] M. S. Vitiello, G. Scalari, B. Williams, and P. De Natale, 'Quantum cascade lasers: 20 years of challenges', *Opt. Express*, **23**(4):5167, 2015.

[6] S. Kumar, 'Recent progress in terahertz quantum cascade lasers', *IEEE J. Sel. Top. Quantum Electron.*, **17**(1):38–47, 2011.

[7] P. Dean, A. Valavanis, J. Keeley, K. Bertling, Y. L. Lim, R. Alhathlool, A. D. Burnett, L. H. Li, S. P. Khanna, D. Indjin, T. Taimre, A. D. Rakić, E. H. Linfield, and A. G. Davies, 'Terahertz imaging using quantum cascade lasers—a review of systems and applications', *J. Phys. D: Appl. Phys.*, **47**(37):374008, 2014.

[8] P. Harrison, 'The nature of the electron distribution functions in quantum cascade lasers', *Appl. Phys. Lett.*, **75**:2800–2802, 1999.

[9] C. Sirtori, P. Kruck, S. Barbieri, P. Collot, J. Nagle, M. Beck, J. Faist, and U. Oesterle, 'GaAs/$Al_xGa_{1-x}As$ quantum cascade lasers', *Appl. Phys. Lett.*, **73**(24):3486–3488, 1998.

[10] V. D. Jovanović, S. Höfling, D. Indjin, N. Vukmirović, Z. Ikonić, P. Harrison, J. P. Reithmaier, and A. Forchel, 'Influence of doping density on electron dynamics in GaAs/AlGaAs quantum cascade lasers', *J. Appl. Phys.*, **99**(10):103106, 2006.

[11] A. Valavanis, L. Lever, C. Evans, Z. Ikonić, and R. Kelsall, 'Theory and design of quantum cascade lasers in (111) n-type Si/SiGe', *Phys. Rev. B*, **78**(3):035420, 2008.

[12] D. Indjin, P. Harrison, R. W. Kelsall, and Z. Ikonic, 'Influence of leakage current on temperature performance of GaAs/AlGaAs quantum cascade lasers', *Appl. Phys. Lett.*, **81**:400–402, 2002.

[13] D. Indjin, S. Tomić, Z. Ikonić, P. Harrison, R. W. Kelsall, V. Milanović, and S. Kućinac, 'Gain-maximised GaAs/AlGaAs quantum cascade laser with digitally graded active region', *Appl. Phys. Lett.*, **81**:2163–2165, 2002.

[14] D. Indjin, P. Harrison, R. W. Kelsall, and Z. Ikonić, 'Self-consistent scattering theory of transport and output characteristics of quantum cascade lasers', *J. Appl. Phys.*, **91**(11):9019–9026, 2002.

[15] D. Indjin, P. Harrison, R. W. Kelsall, and Z. Ikonić, 'Mechanisms of temperature performance degradation in terahertz quantum-cascade lasers', *Appl. Phys. Lett.*, **82**(9):1347–1349, 2003.

[16] D. Indjin, P. Harrison, R. Kelsall, and Z. Ikonic, 'Self-consistent scattering model of carrier dynamics in GaAs–AlGaAs terahertz quantum-cascade lasers', *IEEE Photonics Technol. Lett.*, **15**(1):15–17, 2003.

[17] D. Indjin, Z. Ikonić, V. D. Jovanović, P. Harrison, and R. W. Kelsall, 'Mechanisms of carrier transport and temperature performance evaluation in terahertz quantum cascade lasers', *Semicond. Sci. Technol.*, **19**:S104–S106, 2004.

[18] C. Evans, V. Jovanovic, D. Indjin, Z. Ikonic, and P. Harrison, 'Thermal effects in InGaAs/AlAsSb quantum-cascade lasers', *Optoelectr., IEE Proc.*, **153**(6):287–292, 2006.

[19] C. Evans, D. Indjin, Z. Ikonic, P. Harrison, M. Vitiello, V. Spagnolo, and G. Scamarcio, 'Thermal modeling of terahertz quantum-cascade lasers: comparison of optical waveguides', *IEEE J. Quantum Electron.*, **44**(7):680–685, 2008.

[20] V. Spagnolo, G. Scamarcio, H. Page, and C. Sirtori, 'Simultaneous measurement of the electronic and lattice temperatures in GaAs/AlGaAs quantum-cascade lasers: Influence on the optical performance', *Appl. Phys. Lett.*, **84**:3690, 2004.

[21] V. Spagnolo, G. Scamarcio, H. Page, and C. Sirtori, 'Influence of the band-offset on the electronic temperature of GaAs/Al(Ga)As superlattice quantum cascade lasers', *Semicond. Sci. Technol*, **19**:S110, 2004.

[22] V. D. Jovanović, D. Indjin, N. Vukmirović, Z. Ikonić, P. Harrison, E. H. Linfield, H. Page, X. Marcadet, C. Sirtori, C. Worrall, H. E. Beere, and D. A. Ritchie, 'Mechanisms of dynamic range limitations in GaAs/AlGaAs quantum-cascade lasers: influence of injector doping', *Appl. Phys. Lett.*, **86**:211117, 2005.

[23] S. Höfling, V. D. Jovanović, D. Indjin, J. P. Reithmaier, A. Forchel, Z. Ikonić, N. Vukmirović, P. Harrison, A. Mirčetić, and V. Milanović, 'Dependence of saturation effects on electron confinement and injector doping in GaAs/$Al_{0.45}Ga_{0.55}$As quantum-cascade lasers', *Appl. Phys. Lett.*, **88**(25):251109, 2006.

[24] K. Ohtani, Y. Moriyasu, H. Ohnishi, and H. Ohno, 'Above room-temperature operation of InAs/AlGaSb superlattice quantum cascade lasers emitting at 12μm', *Appl. Phys. Lett.*, **90**(26):261112, 2007.

[25] E. Mujagić, M. Austerer, S. Schartner, M. Nobile, L. K. Hoffmann, W. Schrenk, G. Strasser, M. P. Semtsiv, I. Bayrakli, M. Wienold, and W. T. Masselink, 'Impact of doping on the performance of short-wavelength InP-based quantum-cascade lasers', *J. Appl. Phys.*, **103**(3):033104, 2008.

[26] P. Slingerland, C. Baird, and R. H. Giles, 'Application of multi-subband self-consistent energy balance method to terahertz quantum cascade lasers', *Semicond. Sci. Technol.*, **27**(6):065009, 2012.

[27] D. Indjin, P. Harrison, and R. W. Kelsall, 'Self-consistent scattering theory evaluation of thermionic emission and leakage currents in III-V quantum cascade lasers', *Russ. Acad. Sci. J. Phys.*, **67**:259–261, 2003.

[28] V. D. Jovanović, P. Harrison, Z. Ikonić, and D. Indjin, 'Physical model of quantum-well infrared photodetectors', *J. Appl. Phys.*, **96**:269–272, 2004.

[29] N. Vukmirović, Z. Ikonić, I. Savić, D. Indjin, and P. Harrison, 'A microscopic model of electron transport in quantum dot infrared photodetectors', *J. Appl. Phys.*, **100**:074502, 2006.

[30] N. Vukmirović, D. Indjin, Z. Ikonić, and P. Harrison, 'Origin of detection wavelength tuning in quantum dots-in-a-well infrared photodetectors', *Appl. Phys. Lett.*, **88**:251107, 2006.

[31] S. Barik, H. H. Tan, C. Jagadish, N. Vukmirović, and P. Harrison, 'Selective wavelength tuning of self-assembled InAs quantum dots grown on InP', *Appl. Phys. Lett.*, **88**:193112, 2006.

[32] L. Fu, H. H. Tan, I. McKerracher, J. Wong-Leung, C. Jagadish, N. Vukmirović, and P. Harrison, 'Effects of rapid thermal annealing on device characteristics of InGaAs/GaAs quantum dot infrared photodetectors', *J. Appl. Phys.*, **99**:114517, 2006.

[33] L. Fu, I. McKerracher, H. Tan, C. Jagadish, N. Vukmirović, and P. Harrison, 'Effect of GaP strain compensation layers on rapid thermally annealed InGaAs/ GaAs quantum dot infrared photodetectors grown by metal-organic chemical-vapor deposition', *Appl. Phys. Lett.*, **91**:073515, 2007.

[34] P. Harrison, *Computational Methods in Physics, Chemistry and Mathematical Biology: An Introduction*, John Wiley & Sons, Ltd, Chichester, 2001.

[35] P. Harrison, Z. Ikonić, N. Vukmirović, D. Indjin, R. W. Kelsall, and V. D. Jovanović, 'On the incoherence of quantum transport in semiconductor heterostructure optoelectronic devices', in *Proceedings of the 10th Biennial Baltic Electronics Conference*, 2006, p. 11.

[36] P. Harrison, D. Indjin, I. Savić, Z. Ikonić, C. A. Evans, N. Vukmirović, R. W. Kelsall, J. McTavish, V. D. Jovanović, and V. Milanović, 'On the coherence/incoherence of electron transport in semiconductor heterostructure optoelectronic devices', *Proc. SPIE*, **6909**:690912, 2008.

[37] R. C. Iotti and F. Rossi, 'Carrier thermalisation versus phonon-assisted relaxation in quantum cascade lasers: a Monte Carlo approach', *Appl. Phys. Lett.*, **78**:2902, 2001.

[38] C. Jirauschek, 'Monte Carlo study of carrier-light coupling in terahertz quantum cascade lasers', *Appl. Phys. Lett.*, **96**(1):011103, 2010.

[39] G. Agnew, A. Grier, T. Taimre, Y. L. Lim, M. Nikolić, A. Valavanis, J. Cooper, P. Dean, S. P. Khanna, M. Lachab, E. H. Linfield, A. G. Davies, P. Harrison, Z. Ikonić, D. Indjin, and A. D. Rakić, 'Efficient prediction of terahertz quantum cascade laser dynamics from steady-state simulations', *Appl. Phys. Lett.*, **106**(16):161105, 2015.

[40] H. Luo, S. R. Laframboise, Z. R. Wasilewski, G. C. Aers, H. C. Liu, and J. C. Cao, 'Terahertz quantum-cascade lasers based on a three-well active module', *Appl. Phys. Lett.*, **90**(4):041112, 2007.

[41] K. Blum, *Density Matrix Theory and Applications*, Physics of Atoms and Molecules. Plenum Press, New York, 1996.

[42] T. V. Dinh, A. Valavanis, L. J. M. Lever, Z. Ikonić, and R. W. Kelsall, 'Extended density-matrix model applied to silicon-based terahertz quantum cascade lasers', *Phys. Rev. B*, **85**(23):235427, 2012.

[43] G. Beji, Z. Ikonić, C. A. Evans, D. Indjin, and P. Harrison, 'Coherent transport description of the dual-wavelength ambipolar terahertz quantum cascade laser', *J. Appl. Phys.*, **109**(1):013111, 2011.

[44] H. Callebaut and Q. Hu, 'Importance of coherence for electron transport in terahertz quantum cascade lasers', *J. Appl. Phys.*, **98**(10):104505, 2005.

[45] A. Yariv, C. Lindsey, and U. Sivan, 'Approximate analytic solution for electronic wave functions and energies in coupled quantum wells', *J. Appl. Phys.*, **58**(9):3669–3672, 1985.

[46] H. Callebaut, S. Kumar, B. S. Williams, Q. Hu, and J. L. Reno, 'Importance of electron-impurity scattering for electron transport in terahertz quantum-cascade lasers', *Appl. Phys. Lett.*, **84**(5):645–647, 2004.

[47] M. A. Talukder and C. R. Menyuk, 'Temperature-dependent coherent carrier transport in quantum cascade lasers', *New J. Phys.*, **13**(8):083027, 2011.

[48] I. Savić, N. Vukmirović, Z. Ikonić, D. Indjin, R. Kelsall, P. Harrison, and V. Milanović, 'Density matrix theory of transport and gain in quantum cascade lasers in a magnetic field', *Phys Rev B*, **76**:165310, 2007.

[49] S. V. Gryshchenko, M. V. Klymenko, O. V. Shulika, I. A. Sukhoivanov, and V. V. Lysak, 'Temperature dependence of electron transport in GaN/AlGaN quantum cascade detectors', *Superlatt. Microstruct.*, **52**(4):894–900, 2012.

[50] T. Kubis, C. Yeh, and P. Vogl, 'Non-equilibrium quantum transport theory: current and gain in quantum cascade lasers', *J. Comput. Electron.*, **7**(3):432–435, 2007.

[51] A. Wacker, M. Lindskog, and D. Winge, 'Nonequilibrium Green's function model for simulation of quantum cascade laser devices under operating conditions', *IEEE J. Sel. Top. Quantum Electron.*, **19**(5):1–11, 2013.

13

Optical waveguides

with C. A. Evans

University of Leeds

13.1 Introduction to optical waveguides

Optical waveguides are an integral aspect of optoelectronic devices. Gaining the ability to direct the flow of light, just as copper wires and coaxial cable direct the flow of electrons, enabled a 'quantum leap' to be made in areas such as photonics and communications. Optoelectronic devices such as amplifiers, optical fibres, power splitters, reflectors, directional couplers, polarisers, modulators, converters and frequency shifters rely on the ability to guide light to where it is needed. Optical waveguides also play a large role in the operation of semiconductor laser diodes. The gain medium is placed within a cavity, which is terminated at either end with partially reflecting mirrors. This configuration allows photons to travel back and forth along the cavity, causing more and more stimulated emission until the optical gain overcomes the absorption loss and lasing commences. This behaviour gives rises to standing wave patterns between the mirrors, which are known as *longitudinal* waveguide modes (for more information on longitudinal modes, see [1]). The rate at which stimulated emission occurs from an excited state in a laser is directly proportional to the density of photons at the emission frequency in the laser cavity. To this end, it is beneficial to increase the density of photons, and this is done by confining the photons within the cavity using an optical waveguide. An optical waveguide therefore provides transverse light confinement, preventing beam divergence as it propagates along the cavity, which helps to maintain a high optical intensity and hence high gain along its entire length. A schematic representation of a typical laser waveguide cavity is shown in Fig. 13.1.

Although the major focus of this chapter is the application of optical waveguides to semiconductor lasers, much of the theory is equally applicable to other types of waveguide structures. The basic building block that will be focused upon is the *planar waveguide*. The planar waveguide is spatially inhomogeneous in one direction and assumed to stretch to infinity in the other two. Light confinement is therefore one-dimensional and the light is trapped in the region of highest refractive index. Examples of common types of planar waveguides are shown in Fig. 13.2. For the example on the left, light is confined to the thin film on top of the substrate. The high-index film can be deposited on the low-index substrate

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

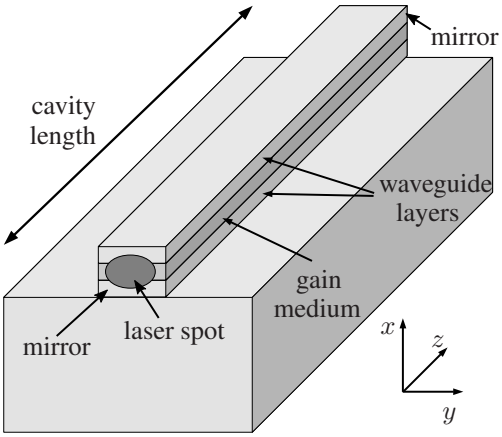


Figure 13.1: Schematic diagram of a laser waveguide cavity

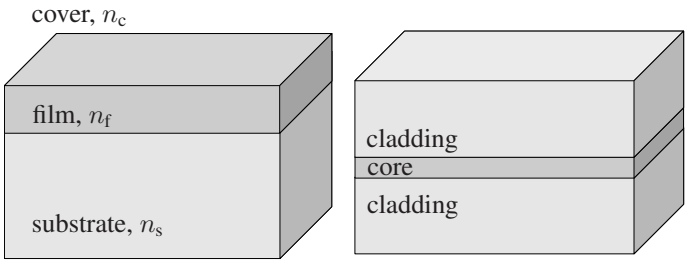


Figure 13.2: Examples of types of planar waveguide: (left) light is confined in the film; (right) light is confined to the core. Darker shading corresponds to higher refractive index

through either physical or chemical means. The cover region is generally formed by the surrounding air. An alternative configuration is shown on the right in Fig. 13.2 in which the high-index core region is surrounded by lower-index cladding. This type of planar waveguide is typical of laser diode waveguides.

Although in earlier chapters the z-direction has referred to the direction of epitaxial growth upon a device substrate, the convention for waveguide analysis is unfortunately different. Here, as shown in Fig. 13.1, z refers to the direction of light propagation along a waveguide (i.e. towards the partially reflecting mirrored ends). The x-direction describes the height above the substrate (i.e. the z-direction in previous chapters!) and the y-direction lies along the width of the waveguide.

It is also possible to construct waveguides that confine light in two dimensions. These types of waveguides are commonly called *ridge* or *channel* waveguides and examples of these are shown in Fig. 13.3. Light is generally confined within the ridge or channel region with vertical confinement provided through the same means as for planar waveguides, i.e. by the lower-index substrate and cover regions. Additionally, horizontal confinement is also

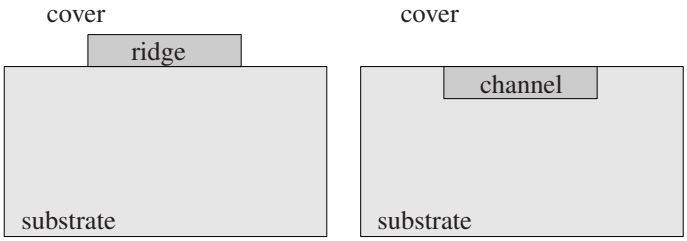


Figure 13.3: Examples of types of two-dimensional waveguides: (left) light is confined in the high-index ridge; (right) light is confined to the high-index channel

provided (e.g. in the case of a ridge waveguide) by the lower-index cover region which surrounds the sides of the ridge. This horizontal confinement is important for laser diodes as it prevents the light beam suffering from divergence as it propagates along the laser cavity. It would therefore seem necessary to study the optical properties of these two-dimensional structures in relation to laser diode waveguides. However, if the width of the ridge or channel is large compared with the wavelength of light (greater than 10λ) in the region, then these types of waveguides can be considered planar waveguides, which simplifies the analysis considerably.

13.2 Optical waveguide analysis

This section will discuss the propagation of light in a planar multilayer waveguide. Due to its simplicity, studying the planar waveguide is a good starting point for understanding more complicated waveguide structures. The theory presented in this section is discussed in more depth in several texts; for instance [1, 2, 3] give a good introduction to optics and photonics aimed both at undergraduates and postgraduates.

Starting from Maxwell's equations, this section derives the equations that are used to describe light propagation in planar waveguides. Once these governing equations have been obtained, a method for solving them is presented, together with an example calculation.

13.2.1 The wave equation

For light that is propagating in an isotropic, non-magnetic (permeability $\mu = \mu_0$), perfect dielectric (conductivity $\sigma = 0$) medium, Maxwell's equations reduce to:

$$\nabla \times \mathbf{E} = -\mu_0 \frac{\partial \mathbf{H}}{\partial t} \tag{13.1}$$

$$\nabla \times \mathbf{H} = \epsilon_0 n^2 \frac{\partial \mathbf{E}}{\partial t} \tag{13.2}$$

where $\mathbf{E}$ and $\mathbf{H}$ are the electric and magnetic fields, ϵ_0 is the permittivity of free space, n is the index of refraction of the material in which the light propagates and $n^2 = \epsilon_\mathrm{r}$ where ϵ_r is the permittivity of the medium. Now, if the medium is inhomogeneous and its refractive index is position-dependent, $n = n(\mathbf{r})$, then the following wave equations for $\mathbf{E}$ and $\mathbf{H}$ can

be derived [2]:

$$\nabla^2\mathbf{E} + \nabla\left(\frac{1}{n^2}\nabla n^2\mathbf{E}\right) - \epsilon_0\mu_0 n^2\frac{\partial^2\mathbf{E}}{\partial t^2} = 0 \tag{13.3}$$

$$\nabla^2\mathbf{H} + \frac{1}{n^2}\nabla n^2 \times (\nabla \times \mathbf{H}) - \epsilon_0\mu_0 n^2\frac{\partial^2\mathbf{H}}{\partial t^2} = 0 \tag{13.4}$$

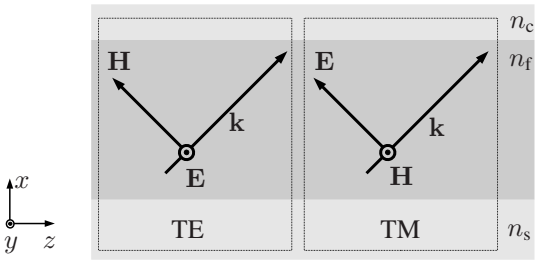


Figure 13.4: Relevant field directions for transverse electric and transverse magnetic propagation in a planar waveguide

In the case of the planar waveguide depicted in Fig 13.2, the refractive index is dependent only upon a single Cartesian coordinate (x) and, assuming that propagation is in the z-direction, the fields do not depend on the y-direction and the solutions to equations (13.3) and (13.4) take the form:

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E_0}(x)\mathrm{e}^{i(\omega t - \beta z)} \tag{13.5}$$

$$\mathbf{H}(\mathbf{r}, t) = \mathbf{H_0}(x)\mathrm{e}^{i(\omega t - \beta z)} \tag{13.6}$$

where $\mathbf{E_0}$ and $\mathbf{H_0}$ are the complex field amplitudes, ω is the angular frequency and β the propagation constant. These two expressions determine the electromagnetic (EM) field for a propagating mode, characterised through its propagation constant. Therefore, for a given planar waveguide (defined by its refractive index profile $n(x)$), the solutions required are the complex field amplitudes and the propagation constants. Since the complex field amplitudes are determined by the propagation constants, the task of solving a waveguide problem becomes one of finding the propagation constants. A method to calculate the propagation constants of a planar waveguide is presented in Section 13.2.2.

As has just been mentioned, the fields do not depend on the y-direction. Setting $\partial/\partial y = 0$ in equations (13.1) and (13.2), Maxwell's equations reduce to:

$$\epsilon_0 n^2\frac{\partial E_x}{\partial t} = -\frac{\partial H_y}{\partial z} \tag{13.7}$$

$$\epsilon_0 n^2\frac{\partial E_y}{\partial t} = \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} \tag{13.8}$$

$$\epsilon_0 n^2\frac{\partial E_z}{\partial t} = \frac{\partial H_y}{\partial x} \tag{13.9}$$

and

$$-\mu_0 \frac{\partial H_x}{\partial t} = -\frac{\partial E_y}{\partial z} \tag{13.10}$$

$$-\mu_0 \frac{\partial H_y}{\partial t} = \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} \tag{13.11}$$

$$-\mu_0 \frac{\partial H_z}{\partial t} = -\frac{\partial E_y}{\partial x} \tag{13.12}$$

This group of equations contains two sets of solutions. Equations (13.8), (13.10) and (13.12) contain terms only in E_y, H_x and H_z, while equations (13.7), (13.9) and (13.11) contain only H_y, E_x and E_z terms. The first set form solutions in which the electric field has no components in the direction of propagation (i.e. only transverse components). These solutions are known as *transverse electric* (TE) modes. Similarly, the second set form solutions in which the magnetic field has no components in the direction of propagation. These are known as *transverse magnetic* (TM) modes.

Transverse electric modes

By substituting the equations for the electric and magnetic fields from equations (13.5) and (13.6) into the above set of solutions for TE modes, the following coupled equations are obtained:

$$H_x = -\frac{\beta}{\omega\mu_0} E_y \tag{13.13}$$

$$H_z = \frac{i}{\omega\mu_0} \frac{\partial E_y}{\partial x} \tag{13.14}$$

$$i\beta H_x + \frac{\partial H_z}{\partial x} = -i\omega\epsilon_0 n(x)^2 E_y \tag{13.15}$$

which link the field components E_y, H_x and H_z. By substituting equations (13.13) and (13.14) into equation (13.15), defining $k_0 = 2\pi/\lambda_0$, where λ_0 is the free-space wavelength, then $\omega = 2\pi c/\lambda_0 = k_0 c$ with c being the speed of light and equal to $(\epsilon_0\mu_0)^{-1/2}$, the following wave equation containing only the E_y component is obtained:

$$\frac{\mathrm{d}^2 E_y(x)}{\mathrm{d}x^2} - \kappa^2 E_y(x) = 0 \tag{13.16}$$

where

$$\kappa = \sqrt{\beta^2 - k_0^2 n^2} \tag{13.17}$$

is the transverse wave vector. This can be solved by applying appropriate boundary conditions at the interfaces of the planar waveguide. From Maxwell's equations it can be shown that, at the interface between two dielectric media, the boundary conditions are such that the tangential components of the electric and magnetic fields must be continuous across the interface. In the case of the geometry in Fig. 13.4, the interfaces are in the y–z plane and so E_y and H_z must be continuous across the interface. From equation (13.14) it can be seen that H_z is proportional to the first derivative of E_y and hence the TE boundary conditions imply that E_y and its first derivative must be continuous across the interfaces.

Transverse magnetic modes

Following the same procedure as carried out above for TE modes but utilising the set of solutions for TM modes, the following coupled equations are obtained linking H_y, E_x and E_z:

$$E_x = \frac{\beta}{\omega\epsilon_0 n(x)^2} H_y \tag{13.18}$$

$$E_z = \frac{-i}{\omega\epsilon_0 n(x)^2} \frac{\partial H_y}{\partial x} \tag{13.19}$$

$$i\beta E_x + \frac{\partial E_z}{\partial x} = i\omega\mu_0 H_y. \tag{13.20}$$

Substituting equations (13.18) and (13.19) into equation (13.20) gives the wave equation for TM modes:

$$\frac{\mathrm{d}^2 H_y(x)}{\mathrm{d}x^2} - \frac{1}{n(x)^2}\frac{\mathrm{d}n(x)^2}{\mathrm{d}x}\frac{\mathrm{d}H_y(x)}{\mathrm{d}x} - \kappa^2 H_y(x) = 0 \tag{13.21}$$

As is also the case for TE modes, appropriate boundary conditions are necessary, and for TM modes the requirement for continuous tangential electric and magnetic field components across an interface means that, for the geometry in Fig. 13.4, E_z and H_y must be continuous. From equation (13.19), it can be seen that E_z is proportional to the first derivative of H_y multiplied by $1/n^2$, and hence the TM boundary conditions require that H_y and $(1/n^2)\mathrm{d}H_y/\mathrm{d}x$ must be continuous across the interfaces.

13.2.2 The transfer matrix method

The wave equations for TE and TM modes have now been derived for planar waveguides, and in this section a transfer matrix method for solving the wave equations will be presented. For simple three-layer planar waveguides, relatively straightforward equations can be derived analytically, in closed form, and solved numerically to give the propagation constant of the modes supported by the waveguide (see, for example, [1]). However, this method is not tractable for more complicated planar waveguide structures, and numerical methods are required instead. A transfer matrix method has already been introduced in Section 2.8 for finding the carrier eigenstates in an multiple quantum well structure, and a similar approach is commonly used in optical waveguide problems owing to its flexibility and computational simplicity [4, 5, 6].

The geometry of a multilayer planar waveguide is shown in Fig. 13.5.

The wave equation for TE modes is given in equation (13.16) and it can be shown to have a general solution of the form:

$$E_i(x) = A_i \exp[-\kappa_i(x - x_{i-1})] + B_i \exp[\kappa_i(x - x_{i-1})] \tag{13.22}$$

which can be thought of a superposition of forward and backward propagating electric fields. $E_i(x)$ is the y-component of the electric field amplitude of the wave propagating in the ith layer, A_i and B_i are the complex field coefficients, $\kappa_i = \sqrt{\beta^2 - k_0^2 n_i^2}$ are the transverse

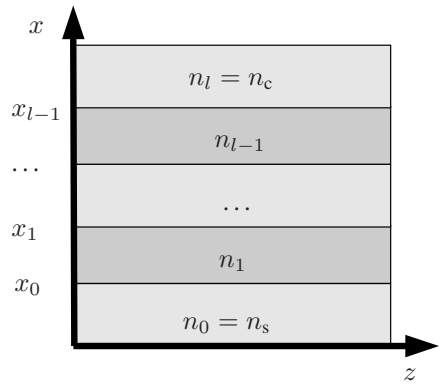


Figure 13.5: Geometry of a multilayer planar waveguide

wave vectors (κ_i can be either real or imaginary) and x_{i-1} is the location of the interface between the ith and $(i-1)$th layer.

The transfer matrix is derived by applying the appropriate boundary conditions at the interface between the layers in the waveguide. As discussed in the previous section, for TE modes, the boundary conditions imply that E_y and its first derivative $\mathrm{d}E_y/\mathrm{d}x$ are continuous across the interfaces. Therefore in the ith layer:

$$E_i(x) = A_i \exp[-\kappa_i(x - x_{i-1})] + B_i \exp[\kappa_i(x - x_{i-1})] \tag{13.23}$$

$$\frac{\mathrm{d}E_i(x)}{\mathrm{d}x} = \kappa_i \left(-A_i \exp[-\kappa_i(x - x_{i-1})] + B_i \exp[\kappa_i(x - x_{i-1})]\right) \tag{13.24}$$

and in the $(i+1)$th layer:

$$E_{i+1}(x) = A_{i+1} \exp[-\kappa_{i+1}(x - x_i)] + B_{i+1} \exp[\kappa_{i+1}(x - x_i)] \tag{13.25}$$

$$\frac{\mathrm{d}E_{i+1}(x)}{\mathrm{d}x} = \kappa_{i+1} \left(-A_{i+1} \exp[-\kappa_{i+1}(x - x_i)] + B_{i+1} \exp[\kappa_{i+1}(x - x_i)]\right) \tag{13.26}$$

At the interface between the ith and $(i+1)$th layers, i.e. at $x = x_i$, the values of equations (13.23) and (13.25) are equal and give the following:

$$A_i \exp[-\kappa_i d_i] + B_i \exp[\kappa_i d_i] = A_{i+1} + B_{i+1} \tag{13.27}$$

where $d_i = x_i - x_{i-1}$ is the thickness of the ith layer. Similarly, the values of equations (13.24) and (13.26) are also equal at $x = x_i$ and give the following:

$$\kappa_i \left(-A_i \exp[-\kappa_i d_i] + B_i \exp[\kappa_i d_i]\right) = \kappa_{i+1} \left(-A_{i+1} + B_{i+1}\right) \tag{13.28}$$

The above equations can be manipulated to give expressions for A_{i+1} and B_{i+1}. Equation (13.27) gives:

$$A_{i+1} = A_i \exp[-\kappa_i d_i] + B_i \exp[\kappa_i d_i] - B_{i+1} \tag{13.29}$$

and equation (13.28) gives:

$$B_{i+1} = A_{i+1} - \frac{\kappa_i}{\kappa_{i+1}} A_i \exp[-\kappa_i d_i] + \frac{\kappa_i}{\kappa_{i+1}} B_i \exp[\kappa_i d_i] \tag{13.30}$$

Inserting these into one another, the following expressions are obtained:

$$A_{i+1} = \frac{1}{2}\left[A_i\left(1 + \frac{\kappa_i}{\kappa_{i+1}}\exp[-\kappa_i d_i]\right) + B_i\left(1 - \frac{\kappa_i}{\kappa_{i+1}}\exp[\kappa_i d_i]\right)\right] \tag{13.31}$$

$$B_{i+1} = \frac{1}{2}\left[A_i\left(1 - \frac{\kappa_i}{\kappa_{i+1}}\exp[-\kappa_i d_i]\right) + B_i\left(1 + \frac{\kappa_i}{\kappa_{i+1}}\exp[\kappa_i d_i]\right)\right] \tag{13.32}$$

The same procedure is can also be carried out for TM modes. In each individual layer, the refractive index is constant and so $\mathrm{d}n(x)^2/\mathrm{d}x = 0$ and the wave equation for TM modes (equation (13.21)) reduces to:

$$\frac{\mathrm{d}^2 H_y(x)}{\mathrm{d}x^2} - \kappa^2 H_y(x) = 0 \tag{13.33}$$

which is equivalent to that for TE modes and has a general solution of the form:

$$H_i(x) = A_i \exp[-\kappa_i(x - x_{i-1})] + B_i \exp[\kappa_i(x - x_{i-1})] \tag{13.34}$$

where $H_i(x)$ is the y-component of the magnetic field amplitude of the wave propagating in the ith layer. It should be noted that the complex coefficients A_i and B_i are *not* the same as for TE modes.

TM boundary conditions require that H_y and $(1/n^2)\mathrm{d}H_y/\mathrm{d}x$ are continuous across the interfaces between the layers. Therefore in the ith layer:

$$H_i(x) = A_i \exp[-\kappa_i(x - x_{i-1})] + B_i \exp[\kappa_i(x - x_{i-1})] \tag{13.35}$$

$$\frac{1}{n^2}\frac{\mathrm{d}H_i(x)}{\mathrm{d}x} = \frac{1}{n_i^2}\left[\kappa_i\left(-A_i \exp[-\kappa_i(x - x_{i-1})] + B_i \exp[\kappa_i(x - x_{i-1})]\right)\right] \tag{13.36}$$

and in the $(i + 1)$th layer:

$$H_{i+1}(x) = A_{i+1} \exp[-\kappa_{i+1}(x - x_i)] + B_{i+1} \exp[\kappa_{i+1}(x - x_i)] \tag{13.37}$$

$$\frac{1}{n^2}\frac{\mathrm{d}H_{i+1}(x)}{\mathrm{d}x} = \frac{1}{n_{i+1}^2}\left[\kappa_{i+1}\left(-A_{i+1} \exp[-\kappa_{i+1}(x - x_i)] + B_{i+1} \exp[\kappa_{i+1}(x - x_i)]\right)\right] \tag{13.38}$$

The same procedure is then followed as for TE modes to obtain the following expressions:

$$A_{i+1} = \frac{1}{2}\left[A_i\left(1 + \frac{n_{i+1}^2}{n_i^2}\frac{\kappa_i}{\kappa_{i+1}}\exp[-\kappa_i d_i]\right) + B_i\left(1 - \frac{n_{i+1}^2}{n_i^2}\frac{\kappa_i}{\kappa_{i+1}}\exp[\kappa_i d_i]\right)\right] \tag{13.39}$$

$$B_{i+1} = \frac{1}{2}\left[A_i\left(1 - \frac{n_{i+1}^2}{n_i^2}\frac{\kappa_i}{\kappa_{i+1}}\exp[-\kappa_i d_i]\right) + B_i\left(1 + \frac{n_{i+1}^2}{n_i^2}\frac{\kappa_i}{\kappa_{i+1}}\exp[\kappa_i d_i]\right)\right] \tag{13.40}$$

Noting the similarity between the above two equations for A_{i+1} and B_{i+1} and equations (13.31) and (13.32), these can be put into matrix form:

$$\begin{pmatrix} A_{i+1} \\ B_{i+1} \end{pmatrix} = \mathsf{Q}_i \begin{pmatrix} A_i \\ B_i \end{pmatrix}$$

where the transfer matrix of the ith layer, Q_i, is given by:

$$\mathsf{Q}_i = \frac{1}{2} \begin{pmatrix} \left[1 + f_i \frac{\kappa_i}{\kappa_{i+1}}\right] \exp[-\kappa_i d_i] & \left[1 - f_i \frac{\kappa_i}{\kappa_{i+1}}\right] \exp[\kappa_i d_i] \\ \left[1 - f_i \frac{\kappa_i}{\kappa_{i+1}}\right] \exp[-\kappa_i d_i] & \left[1 + f_i \frac{\kappa_i}{\kappa_{i+1}}\right] \exp[\kappa_i d_i] \end{pmatrix}$$

For the case of TE modes, f_i=1 and for TM modes, $f_i = n_{i+1}^2/n_i^2$. The transfer matrix Q_i relates the complex field coefficients at the interface between the ith and $(i-1)$th layer, and for a multilayer waveguide with l layers the transfer matrix of the whole waveguide is given by:

$$\mathsf{Q}_{\text{wg}} = \prod_{i=l-1}^{0} \mathsf{Q}_i \tag{13.41}$$

and relates the field coefficients in the cover and the substrate layer, i.e.

$$\begin{pmatrix} A_{\text{c}} \\ B_{\text{c}} \end{pmatrix} = \mathsf{Q}_{\text{wg}} \begin{pmatrix} A_{\text{s}} \\ B_{\text{s}} \end{pmatrix}$$

where the elements of the transfer matrix are denoted by:

$$\mathsf{Q}_{\text{wg}} = \begin{pmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{pmatrix}$$

13.2.3 Guided modes in multilayer waveguides

Now that the transfer matrix that describes the waveguide has been obtained, the propagation constants of the modes supported by the waveguide can be calculated. For the case of *guided* modes, which are generally the most interesting, the light field must decay in the substrate and cover regions; i.e. for the case of the multilayer waveguide geometry in Fig. 13.5, only the outwards decaying real exponentials are permitted in the two outermost semi-infinite layers. From equation (13.17), it can be seen that this is only possible for the case in which the propagation constants of the modes fulfil the condition:

$$(k_0 n_{\text{s}}, k_0 n_{\text{c}}) < \beta \tag{13.42}$$

In the opposite case, no linear combination of oscillatory solutions can have the outwards decaying property. The guided mode condition can be rewritten in terms of the *effective mode index* N, which is equal to β/k_0 and is therefore dimensionless:

$$n_{\text{s}}, n_{\text{c}} < N \tag{13.43}$$

The effective mode index represents the average refractive index that is 'seen' by that particular mode. Other modes also exist, in particular *radiation* or *leaky* modes. These modes

are those in which the light is not confined to the waveguide and leaks into the surrounding cover and substrate regions, causing losses as the light wave propagates along the waveguide. This situation occurs when $n_\mathrm{c} < N < n_\mathrm{s}$ or $N < n_\mathrm{c}$.

According to equations (13.22) and (13.34), the guided mode condition requires that $A_\mathrm{s} = B_\mathrm{c} = 0$. If A_s and B_s are set to zero and unity, respectively, B_c will equal zero when $q_{22} = 0$. This leads to the dispersion relation for the waveguide $q_{22}(\beta) = 0$, which is solved numerically (e.g. using a Newton–Raphson technique) in order to obtain the propagation constants of the waveguide modes.

The only parameters that are left to find are the complex field coefficients A_i and B_i. These are found from the relation:

$$\begin{pmatrix} A_i \\ B_i \end{pmatrix} = \prod_{j=i-1}^{0} \mathsf{Q}_j \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Once A_i and B_i are known, the field profile can be plotted and the waveguide is described completely except for its amplitude. The amplitude of the field is determined by A_i and B_i, but these values are only determined relative to B_s (hence set to unity). Waveguide modes are yet another example of the linear eigenvalue problem, where the absolute field amplitude is determined from only an external condition—in this case the power of a particular mode.

Modes in a planar waveguide

It is now possible to consider an example multilayer planar waveguide. The substrate, film and cover have refractive index values of $n_\mathrm{s} = 3.4$, $n_\mathrm{f} = 3.5$ and $n_\mathrm{c} = 1$, respectively, and the thickness of the film is $5\,\mu\mathrm{m}$. The modes have been calculated for a free-space wavelength of $1.55\,\mu\mathrm{m}$. There exist five guided TE modes in the structure, which can clearly be seen from Fig. 13.6.

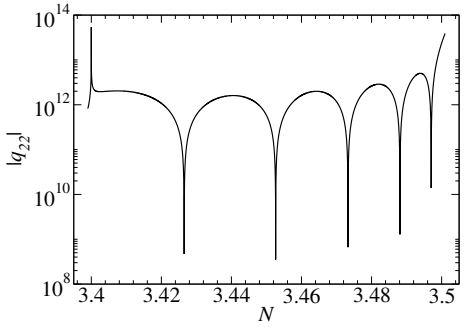


Figure 13.6: Absolute value of q_{22} as a function of effective mode index. Each local minimum represents a TE mode

Since q_{22} is a complex quantity, its absolute value is plotted as a function of effective mode index N. Each minimum in the plot represents a mode; the fact that $|q_{22}|$ does not equal zero

at each mode is entirely down to the numerical accuracy of the zero-finding algorithm. A finer step can be used in the search for zeros, which will obtain smaller values of $|q_{22}|$ but with negligible difference to the final value of N. Modes are generally labelled according to their index $m = 0, 1, 2, \ldots$, starting with the mode with the highest effective mode index. The $m = 0$ mode is known as the fundamental mode. The effective mode indices of both the TE and TM modes supported by the waveguide are listed in Table 13.1.

Table 13.1 Effective mode indices for the TE and TM modes supported by the waveguide

Mode	Index, N	Mode	Index, N
TE_0	3.497026	TM_0	3.496930
TE_1	3.488109	TM_1	3.487731
TE_2	3.473278	TM_2	3.472446
TE_3	3.452630	TM_3	3.451207
TE_4	3.426512	TM_4	3.424463

A plot of the normalised $E_y(x)$ for each TE mode is shown in Fig. 13.7. It can be seen that the electric field and its first derivative are indeed continuous across the interfaces and the field does decay in the cover and substrate regions. Similar results are obtained for the TM modes, except that there exists a discontinuity in the derivative of the magnetic field at the interfaces. The boundary condition for TM modes states that $(1/n^2)\mathrm{d}H_y/\mathrm{d}x$ must be continuous across an interface, and therefore if two adjacent regions have different values of n, then $\mathrm{d}H_y/\mathrm{d}x$ cannot be continuous across the interface in order for the boundary condition to hold.

13.3 Optical properties of materials

In general, the dielectric function of a semiconductor is dispersive and complex:

$$\epsilon(\omega) = \epsilon_{\mathrm{r}}(\omega) + i\epsilon_{\mathrm{i}}(\omega) \tag{13.44}$$

and therefore the refractive index is also complex and frequency dependent:

$$n'(\omega) = n(\omega) + ik(\omega) = \sqrt{\epsilon(\omega)} \tag{13.45}$$

where n is the ordinary (real) refractive index and k is the imaginary part that is also known as the extinction coefficient. The real and imaginary parts of the dielectric constant can be described in terms of the real and imaginary parts of the refractive index:

$$\epsilon_{\mathrm{r}} = n^2 - k^2 \tag{13.46}$$

and

$$\epsilon_{\mathrm{i}} = 2nk \tag{13.47}$$

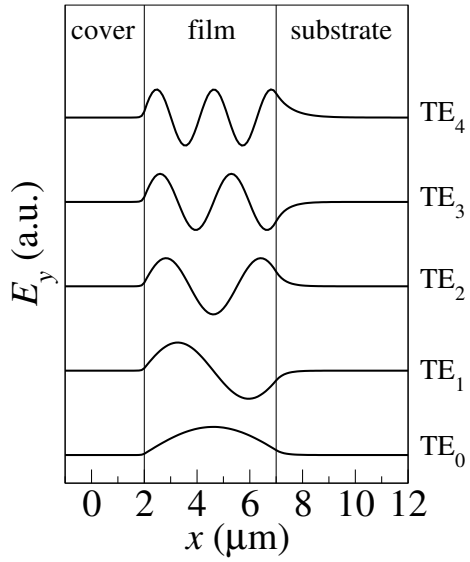


Figure 13.7: Normalised electric field profiles for the TE modes supported in the waveguide. The profiles have been shifted vertically for clarity

Similarly, the real and imaginary parts of the refractive index can be described in terms of the real and imaginary parts of the dielectric constant:

$$n = \left(\frac{\sqrt{\epsilon_{\mathrm{r}}^2 + \epsilon_{\mathrm{i}}^2} + \epsilon_{\mathrm{r}}}{2} \right)^{\frac{1}{2}} \tag{13.48}$$

and

$$k = \left(\frac{\sqrt{\epsilon_{\mathrm{r}}^2 + \epsilon_{\mathrm{i}}^2} - \epsilon_{\mathrm{r}}}{2} \right)^{\frac{1}{2}} \tag{13.49}$$

In general, a material with a purely real refractive index will have zero loss, while a non-zero extinction coefficient will lead to an absorption coefficient:

$$\alpha_k = -2k_0 k = -\frac{4\pi}{\lambda} k \tag{13.50}$$

It should be noted that, in the notation employed in this chapter, a negative value of k (and hence α_k) denotes losses, while a positive value denotes gain (e.g. a laser active medium).

13.3.1 Semiconductors

When modelling waveguides, knowledge of the wavelength dependence of both n and k is crucial. Tabulated values of n and k are available over a wide wavelength range in a host of semiconductors (e.g. [7]). For wavelengths close to and above the band gap, strong interband

absorption occurs, which is characterised by large values of k. For longer wavelengths (i.e. energies below the band edge) these effects are negligible and non-zero values of k are caused by the interaction of photons with phonons. These effects can be modelled very well via a damped harmonic oscillator model [8]:

$$\epsilon(\omega) = \epsilon_\infty \left(1 + \frac{\omega_{\mathrm{LO}}^2 - \omega_{\mathrm{TO}}^2}{\omega_{\mathrm{TO}}^2 - \omega^2 - i\omega\gamma_{\mathrm{ph}}}\right) \tag{13.51}$$

where ϵ_∞ is the high-frequency dielectric constant, ω_{LO} and ω_{TO} are the long-wavelength longitudinal optic (LO) and transverse optic (TO) phonon frequencies respectively, and γ_{ph} is the phonon damping constant. Table 13.2 lists the values used in equation (13.51) for the important III–V semiconductor binaries [7, 8].

Table 13.2 Binary material parameters used in the damped single harmonic oscillator model

Material	ϵ_∞	$\hbar\omega_{\mathrm{LO}}$ (meV)	$\hbar\omega_{\mathrm{TO}}$ (meV)	$\hbar\gamma_{\mathrm{ph}}$ (meV)
GaAs	10.89	36.22	33.32	0.30
AlAs	8.48	49.78	44.86	0.99
InAs	12.25	29.76	27.03	0.50
InP	9.61	42.78	37.65	0.43

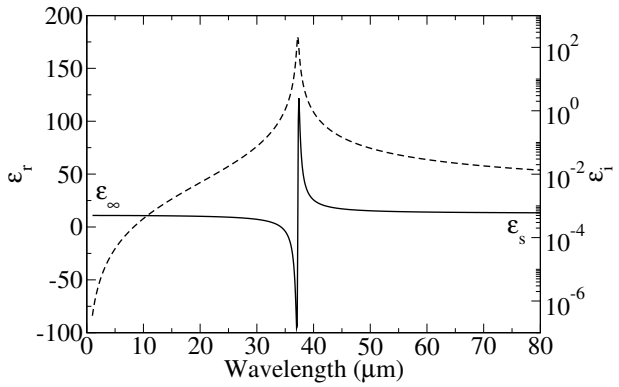


Figure 13.8: The real (solid line) and imaginary (dashed line) parts of the dielectric function of GaAs

The real and imaginary parts of the dielectric function of GaAs are plotted in Fig. 13.8. The dispersion that can be seen is caused by the EM field at these frequencies (called the Reststrahlen region) interacting with fundamental lattice vibrations. This results in the absorption or emission of EM waves due to lattice vibrations being annihilated or created.

It is important to see the difference in the real part of the dielectric constant above and below the Reststrahlen region. On the long-wavelength side of the Reststrahlen region (lower frequency), the refractive index is higher and tends towards the root of the static dielectric constant (ϵ_s), while on the shorter-wavelength side (higher frequency) the refractive index is lower and tends towards the root of the high-frequency dielectric constant (ϵ_∞). The difference in the dielectric constants is due to different polarisation mechanisms that come into play at different frequencies. At frequencies below the Reststrahlen band (far-infrared), the phase velocity is determined by both electronic and ionic polarisation; however, above the Reststrahlen band at optical frequencies, ionic polarisation is too slow to respond to the EM field and hence the phase velocity is determined by only electronic polarisation and is therefore larger. Given that the refractive index can be defined as $n = \sqrt{\epsilon} = c/\nu$ (where ν is the phase velocity), a larger phase velocity results in a lower refractive index. The real part of the dielectric constant is equal to zero when $\omega = \omega_{TO}$.

From equation (13.47) it is apparent that k (and hence α_k) is related directly to ϵ_i, and therefore for wavelengths around the Reststrahlen region extremely strong absorption occurs, preventing laser emission in this wavelength range.

13.3.2 Influence of free carriers

When a semiconductor is doped, equation (13.51) must be modified to include contributions from the free carriers. This contribution is commonly modelled using the classical Drude free-electron model. In this instance the dielectric function (ignoring contributions from phonons) is given by:

$$\epsilon(\omega) = \epsilon_\infty \left(1 - \frac{\omega_p^2}{\omega^2 + i\omega\gamma_{pl}}\right) \tag{13.52}$$

where the plasma frequency is defined as:

$$\omega_p = \sqrt{\frac{N_d q^2}{\epsilon_0 \epsilon_\infty m^*}} \tag{13.53}$$

where N_d is the doping (i.e. free-carrier) density and m^* the effective mass (see Appendix A).

The plasma damping frequency is given by:

$$\gamma_{pl} = \frac{q}{m^* \mu} \tag{13.54}$$

where μ is the carrier mobility. From this equation it is apparent that the plasma damping frequency is actually the inverse of the electron relaxation time τ. Equation (13.52) can be split into real and imaginary parts:

$$\mathrm{Re}[\epsilon(\omega)] = \epsilon_\infty \left(1 - \frac{\omega_p^2}{\omega^2 + \gamma_{pl}^2}\right) \tag{13.55}$$

and

$$\mathrm{Im}[\epsilon(\omega)] = -\frac{\epsilon_\infty \gamma_{pl} \omega_p^2}{\omega(\omega^2 + \gamma_{pl}^2)} \tag{13.56}$$

For the situation when $\omega \gg \gamma_{\mathrm{pl}}$:

$$\mathrm{Re}[\epsilon(\omega)] = \epsilon_\infty \left(1 - \frac{\omega_\mathrm{p}^2}{\omega^2}\right) \tag{13.57}$$

and, if $\omega = \omega_\mathrm{p}$, then $\mathrm{Re}[\epsilon(\omega)] = 0$ and the material is defined as a resonant plasma. When $\mathrm{Re}[\epsilon(\omega)] < 0$, the electron gas in the plasma exhibits metallic behaviour, while, for situations in which $\mathrm{Re}[\epsilon(\omega)] > 0$, the electron gas exhibits dielectric behaviour. Therefore, by tuning the plasma frequency of a semiconductor via a suitable choice of doping density, one can modify the dielectric constant and hence the refractive index. This effect is highlighted in Fig. 13.9, which plots the real part of the refractive index of GaAs as a function of doping density at a wavelength of $9.4\,\mu\mathrm{m}$.

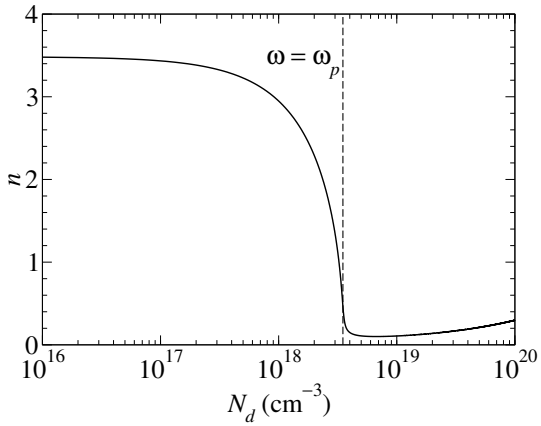


Figure 13.9: Real part of the refractive index of GaAs at various doping densities at a wavelength of $9.4\,\mu\mathrm{m}$. The dashed vertical line indicates the doping density at which the corresponding plasma frequency equals the photon frequency

From the figure it is apparent that, in the limit of low doping, the refractive index approaches the root of ϵ_∞. As the doping is increased and ω_p approaches ω, the refractive index decreases and experiences a strong reduction when $\omega_\mathrm{p} = \omega$.

The Drude model can also be used to calculate the dielectric constant of metals. The plasma frequencies and plasma damping frequencies have been extracted from measurements for a wide range of metals. Some of these are listed in Table 13.3, and ϵ_∞ has been taken to be unity [9].

For frequencies above the plasma frequency, the Drude scattering time is generally long enough such that $\omega \gg \gamma_{\mathrm{pl}}$ and equation (13.56) reduces to:

$$\mathrm{Im}[\epsilon(\omega)] = -\frac{\epsilon_\infty \gamma_{\mathrm{pl}} \omega_\mathrm{p}^2}{\omega^3} \tag{13.58}$$

Table 13.3 Drude parameters for various metals

Metal	$\hbar\omega_{\mathrm{p}}$ (eV)	$\hbar\gamma_{\mathrm{pl}}$ (meV)
Au	9.02	26.67
Ag	9.01	17.98
Cu	7.39	9.08
Al	14.76	81.83

Combining this with equations (13.47) and (13.50) yields:

$$\alpha_k = \frac{k_0 \epsilon_{\mathrm{i}}}{n} = \frac{\epsilon_\infty \gamma_{\mathrm{pl}} \omega_{\mathrm{p}}^2}{4\pi^2 c^3} \lambda^2 \tag{13.59}$$

and in the first instance it is apparent that the loss due to free carriers is proportional to λ^2 which is an important aspect when it comes to long-wavelength waveguides as discussed in Section 13.4.2.

13.3.3 Carrier mobility model

From equation (13.54) it can be seen that the plasma damping frequency is related to the carrier mobility. Since the free-carrier effects discussed in the previous section only come into play at longer wavelengths and in this chapter are only covered in Section 13.4.2 when discussing unipolar lasers, only the electron mobility is required and not the hole mobility. The situation is complicated slightly since the electron mobility itself is a function of doping density due to the influence of ionised impurity scattering. It is therefore useful to have a simple model that accurately describes the influence of doping density on the carrier mobility in a variety of semiconductors. One such model is the empirical Caughey–Thomas mobility model [10, 11]. In this model, the lattice contribution to the electron mobility is given by:

$$\mu^{\mathrm{L}} = \mu_{300}^{\mathrm{L}} \left(\frac{T}{300} \right)^{\gamma_0} \tag{13.60}$$

The total mobility, including the reduction caused by ionised impurity scattering, is given by:

$$\mu = \mu^{\min} + \frac{\mu^{\mathrm{L}} - \mu^{\min}}{1 + \sqrt{\frac{N_{\mathrm{d}}}{N^{\mathrm{ref}}}}} \tag{13.61}$$

where

$$\mu^{\min} = \mu_{300}^{\min} \left(\frac{T}{300} \right)^{\gamma_1} \tag{13.62}$$

and

$$N^{\mathrm{ref}} = N_{300}^{\mathrm{ref}} \left(\frac{T}{300} \right)^{\gamma_2} \tag{13.63}$$

The model parameters are summarised in Table 13.4 for a variety of III–V semiconductors [11] and the resulting mobilities as a function of doping density are plotted

Table 13.4 Parameters used in the Caughey–Thomas mobility model

Material	μ_{300}^{L} ($\mathrm{cm^2\,V^{-1}\,s^{-1}}$)	μ_{300}^{min} ($\mathrm{cm^2\,V^{-1}\,s^{-1}}$)	N_{300}^{ref} ($\mathrm{cm^{-3}}$)	γ_0	γ_1	γ_2
GaAs	8500	800	1×10^{17}	-2.2	-0.9	6.2
AlAs	410	10	1×10^{17}	-2.1	–	–
InAs	32500	11700	4.4×10^{16}	-1.7	-0.33	3.6
InP	5300	1520	6.4×10^{16}	-1.9	2	3.7

in Fig. 13.10. The unknown values can be set to zero in order to remove the temperature dependence. As in the rest of this chapter, T is assumed to be 300 K in all cases. The figure clearly shows the decrease in mobility as the doping density increases due to the increasing influence of ionised impurity scattering.

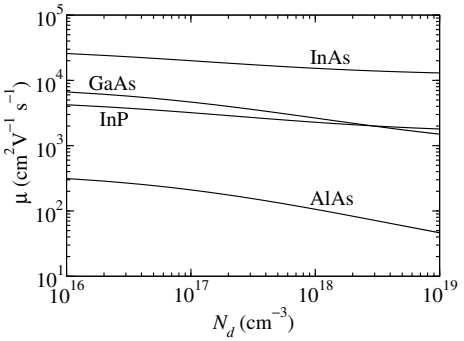


Figure 13.10: Electron mobility of III–V binaries as a function of doping density, as calculated by the Caughey–Thomas mobility model

13.3.4 Influence of doping

The previous sections have introduced models that can be used to calculate the refractive index of various semiconductors as a function of wavelength and doping density. For a given semiconductor at a given wavelength, the phonon contribution to the dielectric constant is calculated using equation (13.51). If the semiconductor is doped, the dielectric constant is modified by the presence of free carriers and is described via the Drude model given in equation (13.52) with the plasma frequency determined by equation (13.53) and the plasma damping frequency by equation (13.54). Both of these parameters depend on the doping density: the plasma frequency has a direct dependence, while the damping frequency depends on doping density indirectly via the mobility, which can be calculated using the model described in Section 13.3.3. It is therefore interesting to study the behaviour of the optical properties of a semiconductor as a function of wavelength and doping density.

Figure 13.11(a) shows a plot of the real part of the refractive index of GaAs as a function of wavelength for different values of doping density. It shows that, irrespective of the doping

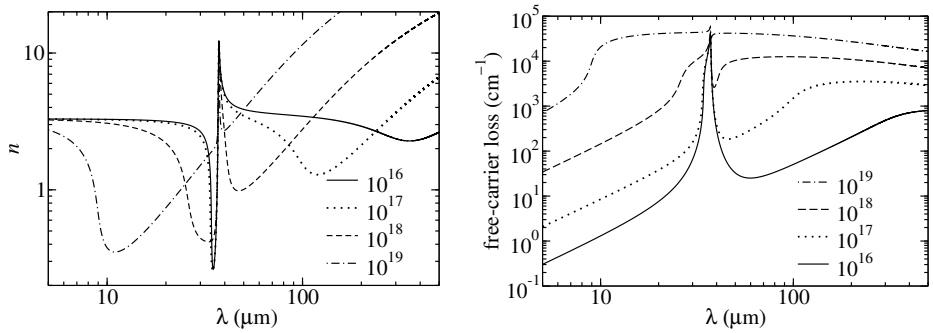


Figure 13.11: (a) Refractive index and (b) free-carrier loss in GaAs as a function of wavelength for various doping densities

density, there is dispersion around the Reststrahlen region caused by the interaction of the light wave with phonons in the material. Each curve also shows a strong reduction in the real part of the refractive index when the wavelength corresponds to the plasma wavelength of the material. The location of this reduction depends on the doping density and occurs at shorter wavelengths (i.e. higher frequencies) for larger doping densities.

Figure 13.11(b) shows a plot of the free-carrier loss of GaAs as a function of wavelength for different values of doping density. It can be seen that the highest value of free-carrier loss occurs in the Reststrahlen region and is practically independent of the doping for densities less than $10^{19}\,\mathrm{cm}^{-3}$. If the contribution from the Reststrahlen region is ignored, the approximation that free-carrier loss is proportional to λ^2 is apparent only until the plasma frequency, when the loss actually begins to fall with increasing wavelength.

13.4 Application to waveguides of laser devices

This section discusses the application of the previously presented waveguide theory and material optical properties to actual laser devices. An example is given of a typical double heterostructure (DH) laser optical waveguide, together with examples of waveguides that are used for lasers emitting at much longer wavelengths.

A laser waveguide is typically characterised by its confinement factor Γ, the fraction of the light field that overlaps with the active medium, and its loss α_{wg} that must be overcome by the gain of the active medium before lasing can commence. Up to now, only planar waveguides with real refractive indices have been considered. From equation (13.50) it can be seen that a material with a complex refractive index will have a corresponding absorption loss, which is the main contribution to the waveguide loss α_{wg}.[1] For waveguides containing complex refractive index materials, the propagation constant β becomes a complex quantity and is

[1] The waveguide loss can be non-zero for purely real refractive index profiles if leaky modes exist in the structure, but this is beyond the scope of this chapter.

given by:

$$\gamma = \beta - i\alpha \tag{13.64}$$

where β has the same meaning as previously and α is the attenuation constant. The method presented in Section 13.2 is still valid, except that β is replaced by γ and, in order to find the solution of the dispersion relation $q_{22}(\gamma)$, a multi-dimension root-finding algorithm (such as a steepest descent method) must be used.

The light intensity I is related to the field through $I \sim |\mathbf{E}|^2$, and for the field described in equation (13.5):

$$I = |E_0 \mathrm{e}^{i(\omega t - \gamma z)}|^2 = |E_0 \mathrm{e}^{i\omega t} \mathrm{e}^{-i\beta z} \mathrm{e}^{-\alpha z}|^2 = |E_0|^2 \mathrm{e}^{-2\alpha z} \tag{13.65}$$

and therefore the loss suffered by the light intensity propagating in the z-direction in the waveguide cavity is given by $\alpha_{\mathrm{wg}} = -2\alpha$.

For a waveguide with loss α_{wg} and cavity length L, the threshold condition is such that a gain of G_0 is required to overcome the losses:

$$R_1 R_2 \exp[2(G_0 - \alpha_{\mathrm{wg}})L] = 1 \tag{13.66}$$

where R_1 and R_2 are the reflectivities of the mirrors at the ends of the laser cavity. If the mirrors are defined by cleaving the facets, the reflectivity is just the Fresnel reflection coefficient, which for a cavity surrounded by air is given by:

$$R_1 = R_2 = \left(\frac{n-1}{n+1}\right)^2 \tag{13.67}$$

where n is the refractive index of the cavity material. The above threshold condition has the meaning that, under steady-state operation, the round trip gain equals unity and can be rearranged to give an expression for the gain at threshold:

$$G_0 = \alpha_{\mathrm{wg}} + \frac{1}{2L} \ln\left(\frac{1}{R_1 R_2}\right) \tag{13.68}$$

in which the second term on the right-hand side represents the loss suffered at the mirrors distributed over the entire cavity length L and is commonly referred to as the mirror loss α_{m}. Now the gain in the above equation is the gain experienced by the laser mode field and, to get it, the actual material gain of the active medium must be multiplied by the confinement factor Γ of the waveguide, as the light field of the laser will only experience gain in the fraction of it that overlaps with the active medium. An expression for the threshold gain required by the device in order for lasing to occur is given by:

$$G_{\mathrm{th}} = \frac{\alpha_{\mathrm{wg}} + \alpha_{\mathrm{m}}}{\Gamma} \tag{13.69}$$

where the confinement factor Γ is commonly defined as the fraction of the light intensity in the waveguide that overlaps with the gain medium:

$$\Gamma = \frac{\int_{\mathrm{active}} |E|^2 \mathrm{d}x}{\int_{-\infty}^{\infty} |E|^2 \mathrm{d}x} \tag{13.70}$$

The confinement factor is a useful quantity that is often used to compare different waveguides and the behaviour of different modes in a single waveguide.

13.4.1 Double heterostructure laser waveguide

For the simple case of an $Al_xGa_{1-x}As$-based DH laser, the active region is sandwiched between two $Al_xGa_{1-x}As$ layers, one doped n-type and the other p-type (for a more in-depth discussion of DH lasers, see [1, 12]). These layers have a larger aluminium fraction (i.e. larger x) and therefore a wider band gap. Since the band gap and refractive index are inversely related to one another in the vast majority of III–V semiconductors, the wider-bandgap layers have a lower refractive index than the laser core and therefore, rather fortuitously, simultaneously act as waveguide cladding layers. This type of waveguide is depicted schematically in Fig. 13.2.

Consider an example DH laser with an active region aluminium content of $x = 0.198$ and cladding layers consisting of $x = 0.491$. The photon energy is typically of the order of $40\,\mathrm{meV}$ above the band gap and, for this particular active region composition, corresponds to an operating wavelength of $\sim$730 nm (1.7 eV) [13]. At wavelengths close to the band gap, strong interband absorption occurs under no-pumping conditions, causing large values of extinction coefficient in the active region material. In the wider-bandgap cladding layers, the absorption is much weaker at the emission wavelength and can generally be ignored. For the AlGaAs compositions in this particular example, the active region has a complex refractive index of $n' = 3.635 - i0.002$, corresponding to an absorption coefficient of $344.3\,\mathrm{cm}^{-1}$ at $730\,\mathrm{nm}$, while the cladding has a purely real refractive index of 3.368 [13]. It has been assumed that the refractive indices do not depend on the doping level or type of electrical conductivity and that free-carrier absorption is negligible, which is reasonable considering the short emission wavelength.

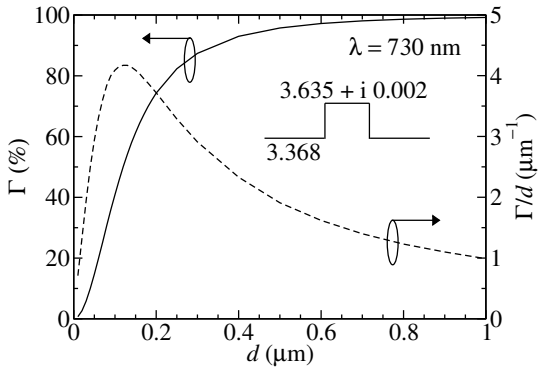


Figure 13.12: Confinement factor (solid line) and ratio of confinement factor to thickness (dashed line) as a function of active region thickness

For this type of symmetric three-layer waveguide, the confinement factor of the waveguide depends on the active region thickness d and the refractive index contrast between the core and cladding layers. For given material compositions, the refractive index contrast is fixed and so Γ can be varied by changing d. In the limit of zero thickness $\Gamma = 0$ and increases with increasing d until it saturates at 100% (see Fig. 13.12). One could be expected to think

that optimum performance is obtained when $\Gamma = 100\%$; however, it has been been shown that the modal gain in a DH laser is proportional to the ratio of Γ/d [14], as a thicker active region requires a larger current density to achieve transparency. For a given current density, the performance of a DH laser with too narrow an active region will be dominated by its small confinement factor, while a DH laser with a wider active region will be limited due to its small gain. Figure 13.12 shows a plot of Γ and the ratio Γ/d as a function of d.

From the figure it can be seen that the optimum thickness of the active region is $d = 0.125\,\mu\mathrm{m}$. Even though the confinement factor Γ is only equal to 52.3%, the ratio of Γ to d is at its peak. At this optimum thickness, the propagation constant of the TE_0 mode is $\gamma = 2.97516 \times 10^6 - i9.46859 \times 10^3\,\mathrm{m}^{-1}$, giving rise to an effective mode index of $N = 3.45663$ and a propagation loss of $\alpha = 189.37\,\mathrm{cm}^{-1}$. For a cavity of length $L = 0.5\,\mathrm{mm}$ and a reflectivity of 0.304, the mirror loss is calculated to be $23.82\,\mathrm{cm}^{-1}$. In an interband laser the absorption and gain are inextricably linked so, as the gain increases (by increasing the current density, for instance), the interband absorption concomitantly decreases. The threshold gain is therefore simply the mirror loss divided by the confinement factor ($g_{\mathrm{th}} = 45.55\,\mathrm{cm}^{-1}$) since, at the point of lasing, the interband absorption and gain are balanced. The electric field distribution (E_y) of the TE_0 mode and the real part of the refractive index of the DH laser are plotted in Fig. 13.13.

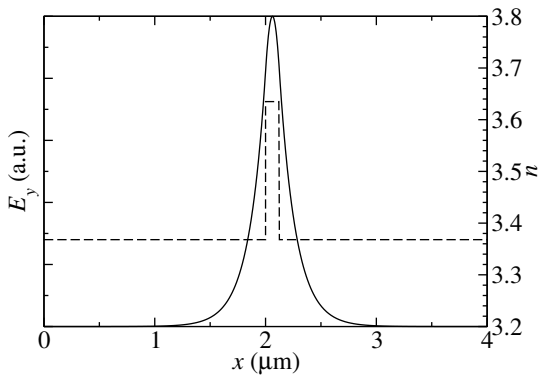


Figure 13.13: Electric field distribution of the TE_0 mode (solid line) and the real part of the refractive index profile (dashed line)

13.4.2 Quantum cascade laser waveguides

Quantum cascade lasers (QCLs) [15] are unipolar lasers that operate on a fundamentally different principle than the DH lasers that were discussed in the previous section. The emission wavelengths of DH lasers are essentially fixed by the band gap of the active medium, while QCLs allow the emission wavelength to be tuned over a wide range from the mid-infrared to the far-infrared (terahertz) region through suitable design of the quantum well and barrier thicknesses. As the emission wavelength is increased, a waveguide that provides

optical confinement through a simple refractive index mismatch becomes impractical since the required thicknesses of the cladding layers become simply too large to be fabricated. Also, free-carrier absorption (equation (13.59)) becomes the dominant loss mechanism at longer wavelengths and can have large values, causing serious implications for device performance. Therefore, as QCL active region designs have evolved to longer and longer wavelengths, so too have the optical waveguides. This section presents typical waveguides that are used for mid- and far-infrared GaAs-based QCLs [16, 17].

Mid-infrared waveguide

As can be seen in Fig. 13.9, by tailoring the plasma frequency of GaAs (and indeed any semiconductor) so that it is close to the emission frequency of the laser through suitable choice of the doping density, one can achieve a significant reduction in the real part of the refractive index. The resulting refractive index contrast that would occur with the laser active region suggests that these highly doped, low-index layers could be used as waveguide cladding layers to provide optical confinement. This is true but, as can be seen from Fig. 13.11(b), the resulting free-carrier absorption at these high doping levels is extremely large and would stop any possibility of laser action. Therefore, in order to circumvent these issues, Sirtori *et al*. placed low-doped layers ($3.5\,\mu\mathrm{m}$ thick) between the active region ($1.6\,\mu\mathrm{m}$) and the highly doped cladding layers ($1\,\mu\mathrm{m}$) in order to separate the optical mode from the lossy highly doped regions [18]. Since the optical confinement is provided by the anomalous dispersion around the plasma frequency, these waveguides are known as double plasmon-enhanced waveguides. Figure 13.14 shows the real part of the refractive index and the free-carrier loss of the waveguide layers calculated at a QCL emission frequency of $\lambda = 9.4\,\mu\mathrm{m}$ using the methods set out in Section 13.3.

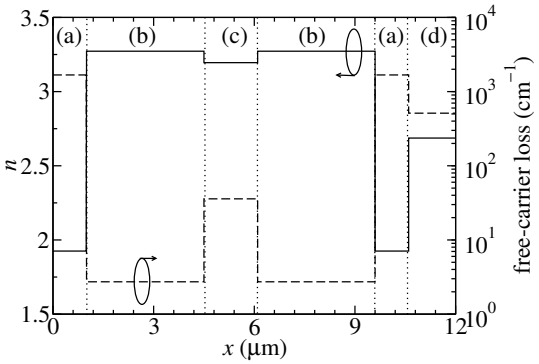


Figure 13.14: Real part of the refractive index profile (solid line) and free-carrier loss (dashed line): (a) highly doped plasmon layer, $N_\mathrm{d} = 6 \times 10^{18}\,\mathrm{cm}^{-3}$, (b) low-doped spacer layer, $N_\mathrm{d} = 4 \times 10^{16}\,\mathrm{cm}^{-3}$, (c) $1.6\,\mu\mathrm{m}$ thick active region, (d) substrate $N_\mathrm{d} = 3 \times 10^{18}\,\mathrm{cm}^{-3}$

Using the field directions depicted in Fig. 13.4, for the case of QCLs, the quantum wells and barriers that form the active region are grown in the x-direction, and, following the

intersubband selection rules, the electric field must be orientated parallel to the growth direction for intersubband transitions to occur. From Fig. 13.4 it can be seen that there is no E_x field component in TE modes, and therefore only TM modes are of interest when it comes to QCL waveguides. When searching for TM mode solutions to the wave equation, H_y is calculated and therefore it seems natural to plot this value. In this case, since generally only E_x is of interest, which is related to H_y via equation (13.20), the mode intensity ($\sim |E_x|^2$) will be plotted in the rest of this section, as this is the quantity that is used in equation (13.70) to compute the confinement factor Γ.

The complex propagation constant of the TM_0 mode of the double plasmon-enhanced waveguide at $9.4\,\mu\mathrm{m}$ is calculated to be $\gamma = 2.14414 \times 10^6 - i8.7040 \times 10^2\,\mathrm{m}^{-1}$, corresponding to an effective mode index of $N = 3.20775$ and a waveguide loss of $\alpha_{\mathrm{wg}} = 17.461\,\mathrm{cm}^{-1}$. The mode intensity of the TM_0 mode is plotted in Fig. 13.15. The

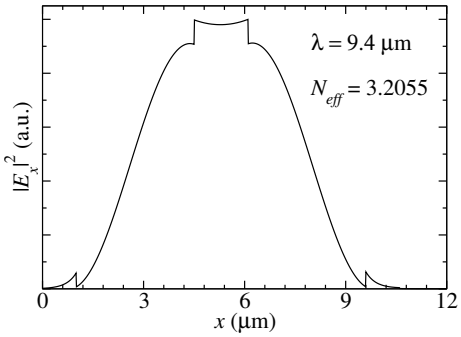


Figure 13.15: Mode intensity of the TM_0 mode in a double plasmon-enhanced QCL waveguide

discontinuities that are observed at the interfaces between layers arise directly from equation (13.20), H_y is set by Maxwell's equations to be continuous across the interfaces, but the different refractive indices of the layers cause the discontinuity in the E_x field.

It has been shown that equation (13.70) is only strictly valid when calculating the confinement factor of TE modes and can lead to inaccuracies when dealing with TM modes [19]. An alternative method to find Γ can be used, as discussed next. This amounts to incorporating the material gain directly into the waveguide calculation, in order to find the threshold gain G_{th}. From equation (13.47) it can be seen that the imaginary part of the permittivity is given by $\epsilon_{\mathrm{i}} = 2nk$. From equation (13.50) it is also known that the extinction coefficient k can be related to the gain/loss of a material via $\alpha_k = -2k_0k$. Therefore, there is a relation that links the imaginary part of the permittivity to the material gain (defined for $\alpha_k < 0$) of the medium:

$$\epsilon_{\mathrm{i}} = \frac{n}{k_0} g_{\mathrm{mat}} \tag{13.71}$$

The previous examples have been worked out in the limit of zero gain, i.e. the absence of lasing. Indeed, the imaginary part of the permittivity of the active region (and hence the

refractive index) is a function of the material gain and, at the onset of lasing, when the gain balances the losses, the medium is transparent. Figure 13.16 shows the waveguide loss α_{wg} as a function of the material gain G_{mat} of the active region. Note that, in the convention used in this chapter, a negative value of waveguide 'loss' denotes gain, while a positive value indicates loss.

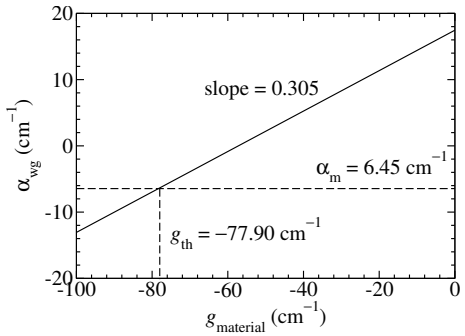


Figure 13.16: Waveguide loss/gain as a function of material gain. The mirror loss α_{m} is indicated by the dashed horizontal line. The intercept of the mirror loss and gain curve gives the threshold gain g_{th}

It can be seen from Fig. 13.16 that, for a material gain of $G_{\mathrm{mat}} = -57.2\,\mathrm{cm}^{-1}$, α_{wg} is zero and the medium appears transparent. In order to extract the threshold gain G_{th}, there must be enough excess gain in order to balance the mirror loss. For a cavity length of $L = 2\,\mathrm{mm}$ and a reflectivity of $R = 0.275$, the mirror loss has a value of $6.45\,\mathrm{cm}^{-1}$. From the figure, it can be seen that, for a material gain of $G_{\mathrm{mat}} = -77.90\,\mathrm{cm}^{-1}$, the waveguide has a gain of $-6.45\,\mathrm{cm}^{-1}$, which balances the mirror loss and therefore $G_{\mathrm{th}} = -77.90\,\mathrm{cm}^{-1}$. Since the permittivity itself is a function of the material gain, so must be the waveguide loss and field distributions. In this essence, the problem is nonlinear but, as can be seen from the figure, for this particular waveguide and range of material gain, the relationship is linear to a very good approximation. In this limit, the slope gives an effective confinement factor and in this case is $\Gamma = 30.5\%$, in very good agreement with the value of $\Gamma = 31.2\%$ that is calculated using equation (13.70). Therefore, for this type of waveguide, equation (13.70) is equally valid for both TE and TM modes.

Far-infrared waveguide

For QCLs with longer emission wavelengths, the double plasmon-enhanced waveguide cannot be used because, as equation (13.59) shows, the free-carrier loss increases according to λ^2, and therefore at these longer wavelengths even the low-doped spacer layers would have extremely unfavourable values of loss, making laser emission difficult. Therefore, alternative waveguide designs based upon surface plasmons [20, 21] have been developed for use in the far-infrared. A plasmon can be thought of as a quantisation of the plasma oscillations

in a free electron gas. A surface plasmon is an electromagnetic wave that is confined to the interface between materials of opposite signs of dielectric constant, i.e. a semiconductor and metal, and exists if $|\epsilon_{\mathrm{m}}| > \epsilon_{\mathrm{s}}$, $\epsilon_{\mathrm{m}} < 0$ and $\epsilon_{\mathrm{s}} > 0$. The loss suffered by a surface plasmon as it propagates along the interface is given by the following [20]:

$$\alpha_{\mathrm{sp}} = \frac{4\pi}{\lambda} \frac{n_{\mathrm{m}} n_{\mathrm{s}}^3}{|k_{\mathrm{m}}^3|} \tag{13.72}$$

where n_{m} and n_{s} are the real parts of the refractive index of the metal and semiconductor, respectively and k_{m} is the extinction coefficient of the metal. The $1/\lambda$ dependence of the loss makes waveguides based upon surface plasmons extremely favourable for long-wavelength devices. This is especially true, since n_{m} and k_{m} themselves depend on λ and so the loss in fact drops faster than $1/\lambda$. The world's first terahertz QCL [17] utilised a so-called semi-insulating surface-plasmon waveguide. In this waveguide, a surface plasmon exists on top of the $\sim 11\,\mu\mathrm{m}$ thick laser active region due to a $200\,\mathrm{nm}$ thick GaAs layer doped to $5 \times 10^{18}\,\mathrm{cm}^{-3}$ followed by a metallic (gold) layer. A highly doped layer ($2 \times 10^{18}\,\mathrm{cm}^{-3}$, $800\,\mathrm{nm}$) is placed under the active region on top of a semi-insulating GaAs substrate. A semi-insulating substrate is used in order to minimise the free-carrier losses at these long wavelengths. At a wavelength of $68\,\mu\mathrm{m}$, the highly doped layer has a complex permittivity of $\epsilon = -94.86 + i41.35$, meaning a surface-plasmon mode is supported at the interface between the highly doped layer and the active region. The overall mode profile of the whole waveguide (Fig. 13.17) can be thought of as a hybridisation of this and the metal–semiconductor plasmon mode.

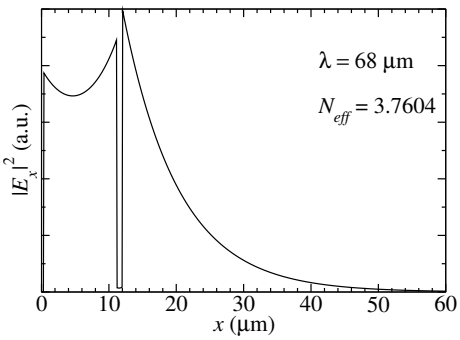


Figure 13.17: Mode intensity of the TM_0 mode in a semi-insulating surface-plasmon QCL waveguide

The propagation constant of the TM_0 mode is calculated to be $\gamma = 3.46740 \times 10^5 + i1.27905 \times 10^3\,\mathrm{m}^{-1}$, corresponding to an effective mode index of $N = 3.76003$ and a waveguide loss of $\alpha_{\mathrm{wg}} = 25.581\,\mathrm{cm}^{-1}$. For a $3\,\mathrm{mm}$ long cavity, the mirror loss is calculated as $3.63\,\mathrm{cm}^{-1}$. The threshold gain and effective confinement factor can be calculated using the same approach as for the double plasmon-enhanced waveguide. A plot of the waveguide loss/gain as a function of material gain is given in Fig. 13.18.

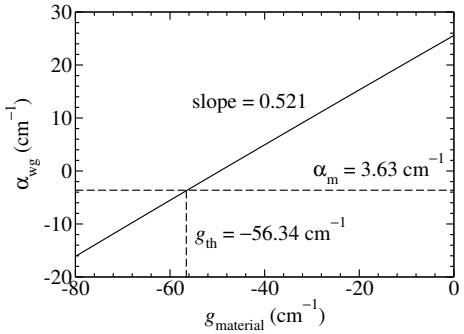


Figure 13.18: Waveguide loss/gain as a function of material gain. The mirror loss α_{m} is indicated by the dashed horizontal line

From the intercept of the mirror loss and gain curves, a threshold gain of $G_{\mathrm{th}} = -56.34\,\mathrm{cm}^{-1}$ is estimated for this waveguide. From the slope of the gain curve an effective confinement factor of $\Gamma = 52.1\%$ is calculated, in good agreement with the value of 49.8% given by equation (13.70).

13.5 Thermal properties of waveguides

So far in this chapter, waveguides have been characterised only in terms of their desired ability to confine radiation within a given volume. However, multilayered waveguide structures also tend to confine *heat*, leading to the existence of localised 'hot spots' within devices. This effect is usually undesirable, since heterostructure devices generally exhibit better performance at low temperatures.

For example, Section 10.9 shows that the rate of carrier scattering caused by phonon absorption increases dramatically as the lattice temperature increases. This leads to an effect, known as *thermal backfilling*, in which carriers can 'climb back up' to higher-energy subbands in intersubband devices. In the case of lasers, for example, this can lead to the lower laser level being refilled by electrons from lower subbands. This destroys the population inversion (see Chapter 12) and prevents the device from lasing at high temperatures.

A second undesirable effect, known as *thermal leakage*, results from the increasing spread in the energy distribution of carriers within a subband as the temperature increases. At high temperatures, carriers no longer cluster near the bottom of a subband, and an increasingly large number of them gain enough energy to scatter into other subbands. In the laser example, this can lead to electrons escaping from the upper laser level into other high-energy states, which again destroys the population inversion.

It is, therefore, important to understand the processes that cause heat to accumulate within devices, in order to design structures that minimise this effect. There are two principal contributions to the temperature within a device. First, the temperature T_{H} of the heat-sink on which the device is mounted defines the minimum temperature within the device: indeed, if the device is unpowered, its internal temperature will eventually settle to that of the heat-sink. Second, the electrical power dissipation IV within the device causes *Joule heating* (or

'resistive' heating), which leads to a temperature increase IVR_{th}, where R_{th} is the *thermal resistance* between the heat-sink and the interior of the device.

At steady state, i.e. if the input power is constant, and the device temperature has been allowed to stabilise over time, the final, steady internal temperature of the device is given by:

$$T_{\text{int}} = T_{\text{H}} + IVR_{\text{th}} \tag{13.73}$$

This expression shows that there are essentially three methods of reducing the internal temperature of a device. The simplest are either to reduce the heat-sink temperature or to reduce the electrical power. However, T_{H} must clearly remain greater than $0\,\text{K}$, and, in practical terms, very low heat-sink temperatures require costly cryogenic cooling systems. Reducing the electrical input power may often result in a trade-off between the internal temperature and another desirable parameter (e.g. the radiative output power of a laser), and may have a lower limiting value, for example, the current through a laser cannot be reduced below the threshold value. The final method of minimising T_{int} is to reduce the thermal resistance, which requires an understanding of the thermal properties of the materials within a device and the mechanisms by which heat is transferred across multilayered structures. The following sections introduce the methods for solving this kind of problem, both in the steady state (as discussed above) and in response to a time-varying input power.

13.6 The heat equation

Fourier's law of heat conduction relates the heat flux $\mathbf{q}$ (i.e. the flow of heat energy) to the gradient in temperature ∇T across a material:

$$\mathbf{q} = -k\nabla T \tag{13.74}$$

where k is the thermal conductivity of the material. The negative sign indicates that heat flows from a hot region to a cold region.

This heat flux leads to changes in the stored energy density at a given location according to:

$$\frac{\partial E}{\partial t} = Q(t) - \nabla \cdot \mathbf{q} \tag{13.75}$$

where $Q(t)$ is the electrical input power per unit volume and $\nabla \cdot \mathbf{q}$ is the net rate of heat transfer away from the location being considered.

The change in energy density can also be expressed in terms of the specific heat capacity of the material c_{p} and its density ρ, such that:

$$\frac{\partial E}{\partial t} = \rho c_{\text{p}} \frac{\partial T}{\partial t} \tag{13.76}$$

Substituting this and equation (13.74) into equation (13.75) yields the heat equation:

$$\rho c_{\text{p}} \frac{\partial T}{\partial t} = \nabla \cdot [k\nabla T] + Q(t) \tag{13.77}$$

Therefore, with knowledge of the material properties ρ, c_{p} and k, and the input power density $Q(t)$ at each point in a device, equation (13.77) may be solved to find the temperature

Table 13.5 Temperature-dependent thermal conductivities

Material	Thermal conductivity, k [$\mathrm{W\,m^{-1}\,K^{-1}}$]	Reference
GaAs	$74500T^{-1.30}$	[8]
AlAs	$225270T^{-1.37}$	[22]
InP	$289000T^{-1.45}$	[23]
Copper	$349 + 14710/T$	[24]
Gold	$337 - T/15.15$	[24]
Indium	$93.9 - 6.96 \times 10^{-2}T + 9.86 \times 10^{-5}T^2$	[24]

at any point (and at any time) within the system. Heat can flow in any direction within a system, and equation (13.77) should be solved in three dimensions to account for this. However, returning to the planar waveguide structures considered earlier in this chapter allows for a simplified analysis. In these systems, it can be assumed that the temperature varies only slowly along the plane of the device, provided that the device width is much greater than the thickness of the layers. As such, the temperature variation is only significant vertically (in the x-direction) through the stack of layers, i.e. $\partial T/\partial y = \partial T/\partial z = 0$. The heat equation therefore simplifies to a one-dimensional problem:

$$\rho c_\mathrm{p} \frac{\partial T}{\partial t} = \frac{\partial}{\partial x}\left[k \frac{\partial T}{\partial x}\right] + Q(t) \tag{13.78}$$

13.7 Material properties

As shown in the previous section, the heat equation can be used to calculate the temperature profile at any given time, provided that the material parameters are known at every point in the system. The densities of materials often remain approximately constant over a wide range of temperatures, and can simply be read from tabulated values (see Appendix A). However, the thermal conductivity and specific heat capacity are themselves highly dependent on temperature and must be calculated at each time-step in a device simulation.

13.7.1 Thermal conductivity

Heat is conducted within bulk semiconductor materials principally through vibrations in the crystal lattice (i.e. phonons), and the thermal conductivity of the material is therefore limited by the speed with which phonons can propagate. This is determined by the scattering of phonons in the lattice, principally through phonon–phonon interactions or scattering from defects in the material. These phonon scattering rates in semiconductors decrease as the temperature reduces, and therefore the thermal conductivity increases at low temperatures. A set of temperature-dependent thermal conductivity functions for commonly used materials (including metals and dielectrics) is shown in Table 13.5, which have been obtained from experimental data.

In the case of alloys (e.g. $Ga_{1-x}Al_xAs$), the thermal conductivity may be determined from the constituent material values using Abele's interpolation scheme, for example:

$$\frac{1}{k_{Ga_{1-x}Al_xAs}} = \frac{x}{k_{AlAs}} + \frac{1-x}{k_{GaAs}} + \frac{x(1-x)}{b} \tag{13.79}$$

where a bowing parameter, $b = 3.33\,\mathrm{W\,m^{-1}\,K^{-1}}$, is used to account for the nonlinear interpolation between the tabulated values [8].

A degree of caution is required, however, when modelling semiconductor heterostructures (such as QCL active regions) since the layer thicknesses are small compared with the bulk phonon wavelengths. In this case, phonons are confined into a set of modes, which are determined by the geometry of the structure, in much the same way as an electron is confined into energy subbands within a heterostructure. As such, it is not physically realistic to treat each of the extremely thin layers in the structure as a separate bulk-like region. Instead, the entire stack of layers can be approximated as a single thermal medium, with a lower thermal conductivity than the bulk materials, owing to the phonon confinement and interfacial scattering effects. The thermal conductivity is also strongly anisotropic in two-dimensional structures, with much lower values across the interfaces than in the plane of material growth. For example, in a GaAs/$Ga_{1-x}Al_xAs$ terahertz QCL, the thermal conductivity has been estimated experimentally as around 75% of the bulk alloy value along the plane [24], but very much lower across the interfaces ($k = 9.6T^{-0.14}\ \mathrm{W\,m^{-1}\,K^{-1}}$ [25]). Strictly, this anisotropic system should be simulated by solving the heat equation in three dimensions (e.g. [25]), but for simplicity only heat transport across the interfaces is considered in the present work.

13.7.2 Specific heat capacity

The specific heat capacity is a measure of the energy required to raise the temperature of a unit quantity (i.e. mass or number of atoms) of a material by $1\,\mathrm{K}$. Since heat energy is stored as phonons within a material, the Debye model may be used to determine the specific heat capacity. Detailed derivations of the model can be found in many fundamental solid-state theory texts (e.g. Kittel [26], Chapter 5), and the result will therefore simply be stated here as:

$$c_\mathrm{m} = 9N_\mathrm{A}k_\mathrm{B}\left(\frac{T}{T_\mathrm{D}}\right)^3 \int_0^{T_\mathrm{D}/T} \frac{x^4\mathrm{e}^x}{(\mathrm{e}^x-1)^2}\,\mathrm{d}x \tag{13.80}$$

This expression gives the *molar* specific heat capacity (i.e. the energy required to raise $1\,\mathrm{mole}$ of atoms by $1\,\mathrm{K}$) as a function of the *Debye temperature*, T_D. Here, $N_\mathrm{A} = 6.022 \times 10^{23}\,\mathrm{mol^{-1}}$ is the Avogadro constant, and $k_\mathrm{B} = 1.38 \times 10^{-23}\,\mathrm{J\,K^{-1}}$ is the Boltzmann constant.

The heat equation (13.77) requires the specific heat capacity for a unit *mass*, rather than a molar quantity, and therefore equation (13.80) must be transformed into mass units. The basic geometric units in the Debye model are individual atomic sites (not primitive molecular units), and therefore the mass of 1 mole of atoms in the material must be found. For elemental materials (e.g. Si), this is simply equal to the molar mass of the material M. However, for compound materials (e.g., GaAs), M is the mass of 1 mole of *molecular units*, and as such it must be divided by the number of atoms per molecule N_atom. Therefore, the specific heat

Table 13.6 Debye temperature and molar mass

Material	Debye temperature, T_D [K]	Molar mass, M [$\mathrm{g\,mol^{-1}}$]	Reference
GaAs	360	144.6	[22]
AlAs	446	101.9	[22]
InP	425	145.8	[27]
Copper	343	63.5	[26]
Gold	165	197.0	[26]
Indium	108	114.8	[26]

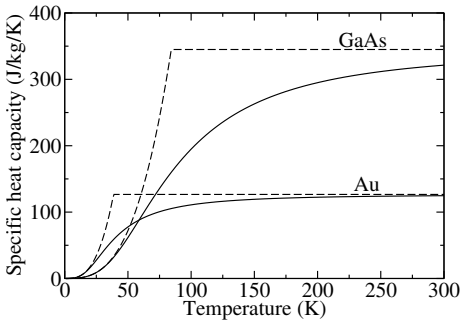


Figure 13.19: Specific heat capacity of GaAs and Au as a function of temperature, using exact solutions of the Debye model (solid lines) and using the low- and high-temperature approximations (dashed lines)

capacity for a unit mass is found using:

$$c_\mathrm{p} = \frac{c_\mathrm{m} N_\mathrm{atom}}{M} \tag{13.81}$$

The Debye model parameters are given for a range of materials in Table 13.6 and its solutions are plotted as a function of temperature for GaAs and gold in Fig. 13.19 (solid lines).

At high temperatures ($T \gg T_\mathrm{D}$), the integrand in equation (13.80) is only computed over a small range ($x \ll 1$) and the exponential term may be replaced by its MacLaurin expansion, $\exp(x) \approx 1 + x$. The Debye model now simplifies to:

$$c_\mathrm{m} = 9 N_\mathrm{A} k_\mathrm{B} \left(\frac{T}{T_\mathrm{D}}\right)^3 \int_0^{T_\mathrm{D}/T} x^2 \,\mathrm{d}x = 3 N_\mathrm{A} k_\mathrm{B} \tag{13.82}$$

and therefore, as the temperature rises above the Debye temperature, the specific heat capacity converges to a constant value, $c_\mathrm{p} = 3 N_\mathrm{A} k_\mathrm{B} N_\mathrm{atom}/M$. The high-temperature specific heat capacity is, therefore, greatest in materials with the lowest average atomic mass. This is apparent in Fig. 13.19, in which the specific heat capacity of Au converges to a lower value than that of GaAs owing to its much greater atomic mass.

The Debye temperature of semiconductors is typically considerably higher than the operating temperatures of heterostructure devices, so the specific heat capacity cannot be

approximated accurately using the high-temperature constant value in this situation. At very low temperatures, the upper limit of the integral in equation (13.80) becomes $T_\mathrm{D}/T \to \infty$, and it can now be solved exactly as:

$$c_\mathrm{m} = \frac{12\pi^4}{5} N_\mathrm{A} k_\mathrm{B} \left(\frac{T}{T_\mathrm{D}}\right)^3 \tag{13.83}$$

Therefore, the specific heat capacity increases as a cubic function of temperature at low temperatures, and is proportional to $1/T_\mathrm{D}^3$. As such, materials with a low Debye temperature tend to exhibit the greatest specific heat capacity at low temperatures. This is seen in Fig. 13.19, in which gold (T_D = 165 K) has a larger specific heat capacity than GaAs (T_D = 360 K) at temperatures up to 58 K. At higher temperatures, the difference in molar mass becomes the dominant factor and the specific heat capacity of GaAs 'overtakes' that of gold.

The low- and high-temperature approximations to the specific heat capacity can be shown to converge with each other at a transition temperature:

$$T_0 = T_\mathrm{D} \left(\frac{5}{4\pi^4}\right)^{1/3} \tag{13.84}$$

Figure 13.19 shows that the approximate models significantly overestimate the specific heat capacity near this transition temperature, and as such it is usually better to solve the Debye model exactly unless the system of interest is operating at temperatures well above or below this transition point.

Finally, it is important to note that the Debye model assumes that the phonon dynamics in the material are constant with respect to temperature. In reality, materials can deviate somewhat from the modelled behaviour as the temperature changes over a large range, but some improvement can be obtained by using a temperature-dependent Debye temperature in the model.

13.8 Finite-difference approximation to the heat equation

As in Sections 3.5 and 4.2, a finite-difference approximation may be used to solve the heat equation numerically. Using the same methods as before, equation (13.78) may be written in the discretised form:

$$T(x, t+\delta t) = a(x,t)T(x-\delta x, t) + b(x,t)T(x,t) + c(x,t)T(x+\delta x, t) + r(x,t)Q(x,t) \tag{13.85}$$

where

$$\begin{aligned} a(x,t) &= \frac{\delta t}{\rho(x,t)c_\mathrm{p}(x,t)\delta x^2}\, k\left(x - \frac{1}{2}\delta x\right) \\ c(x,t) &= \frac{\delta t}{\rho(x,t)c_\mathrm{p}(x,t)\delta x^2}\, k\left(x + \frac{1}{2}\delta x\right) \\ b(x,t) &= 1 - [a(x,t) + c(x,t)] \\ r(x,t) &= \frac{\delta t}{\rho(x,t)c_\mathrm{p}(x,t)} \end{aligned} \tag{13.86}$$

This equation determines the temperature profile through the device at time $t + \delta t$, based on knowledge of the temperature profile at the previous time-step, t. It is, therefore, possible to specify an initial temperature profile for the system, and work forward in time to find the profile at any given time in the future.

The values of the thermal conductivity between the spatial sampling points are chosen so that the heat flux is continuous, i.e.

$$k\left(x+\frac{1}{2}\delta x\right)\left.\frac{\partial T}{\partial x}\right|_{x+\frac{1}{2}\delta x} = k(x)\left.\frac{\partial T}{\partial x}\right|_{x} = k(x+\delta x)\left.\frac{\partial T}{\partial x}\right|_{x+\delta x} \tag{13.87}$$

Now applying the finite-difference approximation to this equation gives:

$$k\left(x+\frac{1}{2}\delta x\right)\frac{T(x+\delta x)-T(x)}{\delta x} = k(x)\frac{T(x+\delta x/2)-T(x)}{\delta x/2} \tag{13.88}$$

and

$$k\left(x+\frac{1}{2}\delta x\right)\frac{T(x+\delta x)-T(x)}{\delta x} = k(x+\delta x)\frac{T(x+\delta x)-T(x+\delta x/2)}{\delta x/2} \tag{13.89}$$

These can be solved simultaneously to give the value of k at the interface as:

$$k\left(x+\frac{1}{2}\delta x\right) = 2\frac{k(x)k(x+\delta x)}{k(x)+k(x+\delta x)} \tag{13.90}$$

A number of different boundary conditions may be specified for the spatial temperature profile. For surfaces of the device attached to the heat-sink, a fixed temperature, $T = T_{\mathrm{H}}$ can be assumed. This is known as the Dirichlet (or isothermal) boundary condition. Conversely, 'Neumann' boundary conditions can be used to describe surfaces that are exposed to a vacuum or a perfect insulator:

$$\frac{\partial T}{\partial x} = 0 \tag{13.91}$$

For the example systems considered here, it will be assumed that one surface of the device is in contact with the heat-sink while the other is in a vacuum. This corresponds to, for example, a QCL attached to a cold finger in a helium cryostat.

13.9 Steady-state solution of the heat equation

If a constant electrical power is used, and the temperature profile of the system is allowed to settle over time, then eventually a steady-state solution to the heat equation will be reached. Under these conditions, $Q(x,t) = Q(x)$ and $\partial T/\partial t = 0$, and therefore equation (13.78) may be rewritten as:

$$\frac{\partial}{\partial x}\left[k\frac{\partial T}{\partial x}\right] = -Q(x) \tag{13.92}$$

It is worth noting that the heat capacity term ρc_{p} no longer appears in the expression. In fact, the heat capacity determines only the rate at which the temperature changes within a structure, rather than its final value. It can be seen that the steady-state solution is determined

only by the power density $Q(x)$, the thermal conductivity $k(x)$ and the boundary conditions of the system.

The time-dependence in the finite-difference approximation can be dropped to yield a simpler expression:

$$d(x)T(x-\delta x)+e(x)T(x)+f(x)T(x+\delta x)=-Q(x) \tag{13.93}$$

where

$$\begin{aligned} d(x) &= \frac{k\left(x-\frac{1}{2}\delta x\right)}{\delta x^2} \\ f(x) &= \frac{k\left(x+\frac{1}{2}\delta x\right)}{\delta x^2} \\ e(x) &= -[d(x)+f(x)] \end{aligned} \tag{13.94}$$

One method of solving this equation is to represent it in matrix form:

$$\mathsf{K}\mathbf{T}=\mathbf{Q} \tag{13.95}$$

where the elements of the vectors $\mathbf{T}$ and $\mathbf{Q}$ give the temperature and power density profiles respectively. K is a real, symmetric tridiagonal matrix containing the thermal conductivity at each point, with its elements given by:

$$\mathsf{K}=\begin{pmatrix} e_1 & f_1 & 0 & \cdots & 0 \\ d_2 & e_2 & f_2 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \cdots & d_{N-1} & e_{N-1} & f_{N-1} \\ 0 & \cdots & 0 & d_N & e_N \end{pmatrix} \tag{13.96}$$

Note that the subscript notation is used here as a shorthand for the spatial location ($e_1 = e(x)$, $e_2 = e(x+\delta x)$ etc).

This kind of symmetrical tridiagonal matrix problem can be solved extremely rapidly using widely available routines in free linear algebra software libraries such as LAPACK [28]. However, it is important to note that the thermal conductivities are themselves temperature-dependent, and as such equation (13.96) is really a nonlinear equation, where all the elements of the K matrix depend on the temperature:

$$\mathsf{K}(\mathbf{T})\ \mathbf{T}=\mathbf{Q} \tag{13.97}$$

As such, it is necessary to solve this iteratively, by selecting an initial estimate of the temperature profile (perhaps using the heat-sink temperature throughout), and then repeatedly solving the equation to find a better estimate of the temperature profile (and hence the thermal conductivities) until the solution converges.

The steady-state temperature of the terahertz QCL described in Section 13.4.2 was simulated. A substrate thickness of $200\,\mu\mathrm{m}$ and a $0.18 \times 3.1\ \mathrm{mm}^2$ waveguide area was used in this structure [17]. The very thin highly doped layers were omitted from the model, since

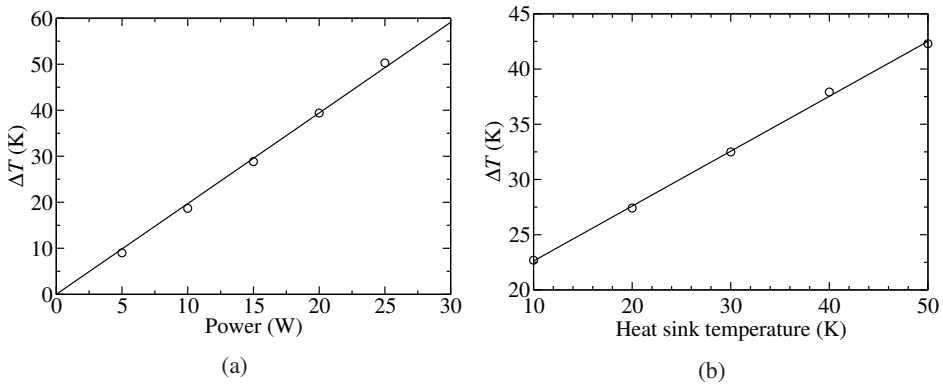


Figure 13.20: Temperature rise in a terahertz QCL active region under continuous bias, as a function of (a) the input power, at a fixed heat-sink temperature of $20\,\mathrm{K}$ and (b) the heat-sink temperature at a fixed input power of $18\,\mathrm{W}$

their total thermal resistance is negligible. The average temperature increase within the active region is shown in Fig 13.20(a) as a function of the electrical input power, using a fixed heat-sink temperature of $20\,\mathrm{K}$. For example, with a steady-state input power of $\sim 18\,\mathrm{W}$, the average temperature of the active region is found to be $35\,\mathrm{K}$ higher than the heat-sink. The temperature difference is shown to rise almost linearly, as expected from equation (13.73), with a thermal resistance of $1.97\,\mathrm{K\,W^{-1}}$. The slight nonlinearity of the temperature rise is caused by the temperature-dependence in the thermal conductivity. Indeed, this effect is more apparent in Fig. 13.20(b), in which the temperature rise is plotted against the heat-sink temperature using a fixed $18\,\mathrm{W}$ input power. For a constant thermal resistance, the plot would show a zero gradient. However, the temperature rise is shown to be related approximately linearly to the heat-sink temperature (at this constant input power), according to the function $\Delta T(P = 18\,\mathrm{K}) = 17.6 + 0.5T_{\mathrm{H}}$. Thus, the net temperature of the terahertz QCL is increased by both the direct contribution of the heat-sink temperature and the increase in thermal resistance. A further important consideration (which has not been modelled here) is the increase in laser threshold current at high temperatures, and hence the requirement for a larger input power. In practice, this leads to an even greater rise in the net temperature of the active region as the heat-sink temperature increases.

The temperature profile of the QCL is shown in Fig. 13.21. The main thermal 'bottleneck' (i.e. the region over which the greatest temperature difference occurs) is within the active region of the device, owing to its much lower thermal conductivity, compared with the bulk GaAs substrate. Even for a large input power of $30\,\mathrm{W}$, the temperature drop across the substrate is only $7\,\mathrm{K}$, compared with $43\,\mathrm{K}$ across the active region. As such, it is tempting to assume that the active region temperature could be reduced (and hence the thermal performance of the laser improved) by reducing the active region thickness. A number of experimental studies of thin active regions have been performed [29, 30]. However, the *optical* performance of waveguides is also reduced significantly as the thickness decreases, with poorer modal confinement and higher waveguide loss. In practice, therefore, thin

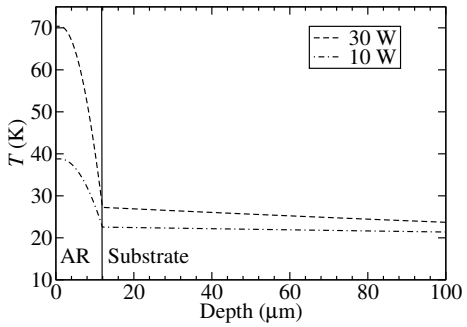


Figure 13.21: Steady-state vertical temperature profile of a terahertz QCL with a constant heat-sink temperature of 20 K. 'AR' denotes the active region of the QCL. Results are shown for constant input powers of 10 and 30 W

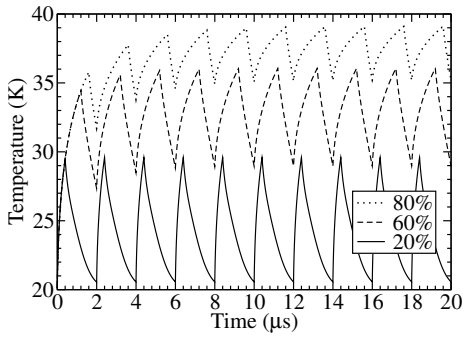


Figure 13.22: Temperature variation at the centre of a QCL active region, in response to a 20 W pulsed input power source at a repetition rate of 500 kHz. A heat sink temperature of 20 K was used. Results are shown for duty cycles of the input pulse of 20%, 60% and 80%

terahertz QCL active regions tend to require higher threshold currents and lower heat-sink temperatures.

13.10 Time-resolved solution

Semiconductor devices are often driven by a pulsed or oscillating power source, and the temperature within the system changes over time without reaching a steady-state value. As such, it is necessary to solve equation (13.78) in the time domain rather than looking for a steady-state solution as in the previous section. This can again be achieved by assuming an initial temperature profile and then solving the finite-difference equation (13.85) at each time-step. As noted in Section 4.5, the forward-time central-space formulation used in equation (13.85) only yields a stable numerical solution if the time and space intervals satisfy

the condition [31]:

$$\frac{k\delta t}{\rho c_{\mathrm{p}}\delta z^2} \leq \frac{1}{2} \tag{13.98}$$

As such, very short time intervals (and hence long simulation times) may be needed for reliable simulation of systems with thin layers. As with the diffusion calculations in Section 4.5, however, more complex numerical techniques can be used to provide an unconditionally stable solution.

Figure 13.22 shows the time-resolved solution for the terahertz QCL discussed in Section 13.9 when driven by a 500 kHz pulsed 20 W power source, and using a fixed heat-sink temperature of 20 K. The figure shows the average active-region temperature for driving signals with a range of different duty cycles (i.e. the fraction of time the QCL is switched on). In all cases, the temperature of the QCL can be seen to alternately increase and decrease by 5–10 K, in response to the input power being switched on and off. Since the pulses are very short, there is insufficient time for the QCL to reach a steady temperature during an individual pulse. The rise and fall times of the individual pulses are on the order of a few microseconds, in good agreement with the time-scales observed experimentally in photoluminescence studies of terahertz QCLs [32].

When the duty cycle is low, the QCL is switched off for sufficient time between each pulse for its temperature to return almost to that of the heat-sink. However, at high duty cycles three important effects are apparent in Fig. 13.22. First, the initial pulse can be seen to lead to a much larger temperature rise, owing to the longer duration of the input power pulse. Second, since the time between the pulses (i.e. with the power supply switched off) is reduced, there is insufficient time for the QCL to cool fully to the heat-sink temperature. Therefore, the temperature continues to rise with each subsequent pulse. This pulse-to-pulse heating effect eventually leads to a quasi-stable condition, in which the individual pulses finally settle into an oscillation (similar to the low-duty-cycle case), but offset to a baseline temperature significantly higher than the heat-sink. This stabilisation is observed over a much longer time-scale (tens or hundreds of microseconds) than the heating of an individual pulse, which can be understood by considering the mechanism of heat flow within the entire device. During each pulse, heat energy is injected into the active region by the electrical power supply, and afterwards it diffuses out of the active region and into the substrate. Since the active region is thin, its heat capacity is very low, and the temperature rises rapidly within each individual pulse. However, the much thicker substrate has a large heat capacity and this limits the speed with which heat diffuses between the two regions. Furthermore, the temperature of the substrate itself rises very slowly compared with that of the active region, and this dictates the pulse-to-pulse heating time. Experimental studies of this effect have shown that pulse-to-pulse heating can take several hundred microseconds to stabilise in terahertz QCLs [33].

The final significant effect visible in Fig. 13.22 is that the final quasi-steady oscillations are slower and smaller when high duty cycles are used. This is caused by the temperature-dependence of the specific heat capacity of the active region. The high-duty-cycle operation initially drives the QCL to a much higher baseline temperature, as discussed above, and the Debye model predicts that the heat capacity will increase as approximately a cubic function of the temperature rise. This rise in heat capacity, therefore, results in the temperature changing much more slowly.

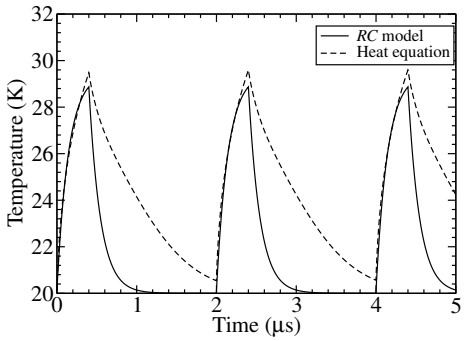


Figure 13.23: Comparison between time-resolved simulations of a QCL using a solution of the heat equation (as in Fig. 13.22) and using a first-order RC model with $R = 0.473\,\mathrm{K\,W^{-1}}$ and $C = 0.304\,\mu\mathrm{J\,K^{-1}}$

13.11 Simplified RC thermal models

The methods described in the sections above allow the analysis of heat conduction through multilayered structures from first principles (i.e. from knowledge of the material properties of the individual layers). While this is useful for determining the behaviour of arbitrary structures, the time-resolved calculations can become relatively time-consuming, particularly for long structures containing many thin layers. Often, the subject of interest is the thermal response of just one part of a single *known structure* in response to an arbitrary electrical signal. In this case, the spatial variation of temperature is no longer required, and instead a very approximate model of the device can be obtained in terms of simplified 'lumped' thermal elements. This is analogous to the study of classical electronic systems, in which a component is considered in terms of a single characteristic property (e.g. the resistance or capacitance), which is more convenient than considering its internal structure in detail.

The *steady-state* behaviour of a complete system has already been considered in terms of a thermal resistance:

$$\Delta T = P R_{\mathrm{th}} \tag{13.99}$$

which relates the flow of thermal energy $P = Q\Omega$ through a volume Ω to the change in temperature $\Delta T = T_{\mathrm{int}} - T_{\mathrm{H}}$ across the structure. This is analogous to the behaviour of an electrical resistor, which relates the flow of current linearly to the change in voltage across the structure. The thermal resistance for a purely resistive structure can be obtained from equation (13.78) by setting the time-dependence $\partial T/\partial t$ to zero to obtain:

$$\frac{P}{\Omega} = -\frac{\partial}{\partial x}\left[k\frac{\partial T}{\partial x}\right] \tag{13.100}$$

This differential equation can then be solved to find the temperature drop across the system as a function of P, and hence the thermal resistance.

This resistive model provides a simplified means of determining the steady-state temperature of a system, but also has the advantages of following the same laws as electrical

resistance. As such, thermal resistances add in series:

$$R_{\text{total}} = R_1 + R_2 + R_3 + \ldots \tag{13.101}$$

This allows a composite system to be modelled very rapidly with no need for solving the heat equation. For example, if a device has thermal resistance $R_{\text{d}} = 10\,\text{K}\,\text{W}^{-1}$, its packaging has thermal resistance $R_{\text{p}} = 1\,\text{K}\,\text{W}^{-1}$ and it is affixed to a heat-sink of temperature $T_{\text{H}} = 10\,\text{K}$ by a layer of indium foil with thermal resistance $R_{\text{i}} = 1\,\text{K}\,\text{W}^{-1}$, then the total thermal resistance between the heat-sink and the top of the device is $R = R_{\text{d}} + R_{\text{p}} + R_{\text{i}} = 12\,\text{K}\,\text{W}^{-1}$. The temperature at the top of the device is then simply $T = T_{\text{H}} + RP = 10 + 12P$.

This purely resistive model does not represent the complete picture, however, since it does not include the gradual heating and cooling effects caused by heat storage, which were discussed in the previous sections. In this case, the ability of a structure to store heat is represented by its total heat capacity, $C = \Omega\rho c_{\text{p}}$, where Ω is its volume. For a pure storage element (i.e. with no thermal conductivity), the spatially varying term in equation (13.78) vanishes, giving:

$$P(t) = C\frac{\partial[\Delta T(t)]}{\partial t} \tag{13.102}$$

This is analogous to a capacitor acting as a store of charge. The temperature difference is then given by the integral of the thermal power over time. As such, this equation shows that in the absence of any thermal conductivity, the temperature rises indefinitely over time when a power source is applied.

Clearly, real thermal systems will exhibit both a degree of heat storage and thermal conductivity, and as such the simplest model will contain a thermal capacitance in parallel with a thermal resistance. This combined lumped-component model can be analysed either in the time domain (by solving coupled differential equations) or in the frequency domain (using techniques taken from electronic circuit theory). In the latter case, the thermal components can be represented by equivalent thermal impedances, $Z(s) = \Delta T(s)/P(s)$, by taking Laplace transforms of equations (13.99) and (13.102) to obtain:

$$Z_R(s) = R_{\text{th}} \tag{13.103}$$

$$Z_C(s) = \frac{1}{sC} \tag{13.104}$$

for the thermal resistance and capacitance respectively, where s is the complex frequency in the Laplace domain. The combined impedance for the parallel thermal resistance and capacitance is then given by:

$$Z(s) = \frac{1}{1/R + sC} = \frac{1/C}{s + 1/(RC)} \tag{13.105}$$

This approach can be applied to any combination of thermal components to generate an effective impedance for coupled systems.

Now this combined impedance can be applied to an arbitrary input power to obtain the temperature difference, according to $\Delta T(s) = Z(s)P(s)$. For example, a step in power of

size P_0 has a Laplace transform of P_0/s, and this yields a temperature difference:

$$\Delta T(s) = \frac{P_0}{s}\,\frac{1/C}{s+1/(RC)} = P_0\left(\frac{R}{s} - \frac{R}{s+1/(RC)}\right) \tag{13.106}$$

Transforming this back to the time domain gives:

$$\Delta T(t) = P_0 R\left(1 - \mathrm{e}^{-t/(RC)}\right) \tag{13.107}$$

Therefore, a step in the input power yields a thermal 'charging' function, in which the temperature rises over time to a final value of $\Delta T = P_0 R$, with time constant RC. This is clearly similar to the behaviour obtained by solving the heat equation from first principles in Section 13.10, but can be solved immediately for a wide range of input power functions without directly solving the differential equations. As such, this first-order RC model is useful for quickly analysing a known system, for example, to estimate the temperature rise associated with changing the pulse width or repetition rate in a laser.

To illustrate the RC modelling approach (and its limitations), equation (13.107) was fitted to the rising edge of the first step in the 20% duty-cycle pulse train of Fig. 13.22, yielding the values $R = 0.473\,\mathrm{K\,W^{-1}}$ and $C = 0.304\,\mu\mathrm{J\,K^{-1}}$. These values were then used in an RC simulation of the complete pulse train. The results are compared with those of the full heat-equation model in Fig. 13.23. The shape and amplitude of the rising edges of the thermal pulses are reproduced quite accurately. However, the first-order model only contains a single time constant, and the fall time of the pulses is underestimated significantly. Clearly, this model is also incapable of simulating the much slower pulse-to-pulse heating effects. The shape of the peaks of the pulses is also slightly inaccurate, since the temperature-dependence of the material properties is ignored.

Although the overall accuracy of this simple model is clearly quite poor, this approach is still widely used owing to its mathematical simplicity and speed of calculation. More sophisticated higher-order models can be used to improve the accuracy, in which multiple RC components are connected together in a more complex thermal circuit. The same circuit-theory approaches as above can then be used to determine the temperature variation. Furthermore, these higher-order models allow a degree of internal information to be determined. For example, by considering a device and its packaging as separate thermal elements, the temperature drop across each component can be analysed independently. Even more sophisticated models can be derived, in which the temperature-dependence of the thermal resistance and heat capacity are included, although this leads to nonlinear circuit equations, which usually need to be solved iteratively.

Exercises

(1). Derive discrete forms of equations (13.16) and (13.21) using the finite-difference approximation. Express the discretised equations as matrix eigenvalue problems using the method described in Section 3.9. Explain the advantages of this approach compared with a transfer matrix method for finding guided modes.

(2). Consider a perfect planar waveguide that consists of a film with refractive index n and thickness L, surrounded by impenetrable films such that the field strength is zero outside the film.

(a) Starting from equation (13.16), derive an analytical expression for the electric field and effective mode index for the mth TE and TM modes in the structure.

(b) Use this analytical method to estimate the first five mode indices for the structure in Fig. 13.7 and comment on the level of agreement with the values in Table 13.1.

(c) Given that a guided mode must have a real-valued index, determine an expression for the maximum number of modes that can exist in the perfect dielectric waveguide. Again, comment on the level of agreement with the five modes found in Table 13.1. Find the minimum thickness for a perfect planar waveguide to support at least one mode.

(3). Starting from equation (13.69), derive an analytical expression that relates the threshold current for a laser to the width, length and thermal resistance of the waveguide ridge. In this derivation, assume that the gain can be related linearly to the current by $G = gJ$, and that the gain coefficient reduces linearly with respect to temperature such that $g = g_0 - \mu\Delta T$, where μ is a scaling constant. Determine expressions for the threshold current in the limits of either very large or small device lengths, and for the maximum length of cavity that could be used. Explain the limit on cavity length qualitatively, and discuss whether this is physically realistic.

(4). Starting from equation (13.78), derive the finite-difference approximation to the heat equation, as in equation (13.85).

(5). Consider a volume of a homogeneous material, with a temperature-independent thermal conductivity k, being heated by a constant power $Q(x,t) = Q_0$. Let one side of the material $x = 0$ be attached to a heat-sink with temperature T_H, and the opposite side $x = L$ be exposed to a vacuum.

(a) Find an analytical expression for the steady-state temperature profile $T(x)$ within the material by solving equation (13.92).

(b) Find the relationship between the temperature difference across the structure and the input power. Comment on the influence of the factors that determine its value and discuss whether it is reasonable to describe this relationship as a thermal resistance.

(c) If the structure is a GaAs cuboid, with a thickness of $100\,\mu\mathrm{m}$ (measured from the heat-sink to the top surface), a length of $4\,\mathrm{mm}$ and a width of $200\,\mu\mathrm{m}$, being driven by a constant current of $2\,\mathrm{A}$ and a voltage of $10\,\mathrm{V}$, and with a heat-sink temperature of $100\,\mathrm{K}$, estimate the temperature at the hottest point in the structure.

(6). A first-order thermal system consists of a heat capacity C and a thermal resistance, R in parallel, as described by equation (13.105). Find an analytical expression for the variation in temperature over time when the device is driven by a sinusoidally varying input power source $P(t) = P_0[1 + \sin(\omega t)]$, and find its steady-state value by letting $t \gg RC$. Show analytically that the system exhibits a low-pass frequency response, i.e. the steady-state temperature oscillations decrease in amplitude as the frequency of the input power oscillations increases (when $\omega \gg 1/RC$).

References

[1] B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley Series in Pure and Applied Optics. John Wiley & Sons, Inc., New York, 1991.

[2] G. Lifante, *Integrated Photonics: Fundamentals*, John Wiley & Sons, Inc., Hoboken, NJ, 2003.

[3] F. Graham-Smith, T. A. King, and D. Wilkins, *Optics and Photonics: an Introduction*, John Wiley & Sons, Ltd., Chichester, 2nd edition, 2007.

[4] E. Anemogiannis and E. N. Glytsis, 'Multilayer waveguides: efficient numerical analysis of general structures', *J. Lightwave Technol.*, **10**(10):1344–1351, 1992.

[5] J. Stiens, R. Vounckx, and I. Veretennicoff, 'Slab plasmon polaritons and waveguide modes in four-layer resonant semiconductor waveguides', *J. Appl. Phys.*, **81**(1):1–10, 1997.

[6] E. Anemogiannis, E. N. Glytsis, and T. Gaylord, 'Determination of guided and leaky modes in lossless and lossy planar multilayer optical waveguides: reflection pole method and wavevector density method', *J. Lightwave Technol.*, **17**(5):929–941, 1999.

[7] E. D. Palik, Ed., *Handbook of Optical Constants of Solids*, Academic Press, San Diego, CA, 1998.

[8] S. Adachi, *GaAs and Related Materials: Bulk Semiconducting and Superlattice Properties*, World Scientific, Singapore, 1994.

[9] M. A. Ordal, R. J. Bell, J. R. W. Alexander, L. L. Long, and M. R. Querry, 'Optical properties of fourteen metals in the infrared and far infrared: Al, Co, Cu, Au, Fe, Pb, Mo, Ni, Pd, Pt, Ag, Ti, V, and W', *Appl. Opt.*, **24**(24):4493, 1985.

[10] D. Caughey and R. Thomas, 'Carrier mobilities in silicon empirically related to doping and field', *Proc. IEEE*, **55**(12):2192–2193, 1967.

[11] V. Palankovski, *Simulation of Heterojunction Bipolar Transistors*, Ph.D. thesis, Technische Universität Wien, 2000.

[12] S. M. Sze, *Physics of Semiconductor Devices*, John Wiley & Sons, Inc., New York, Second edition, 1981.

[13] O. J. Glembocki and K. Takarabe, 'Aluminum gallium arsenide ($Al_xGa_{1-x}As$),' in *Handbook of Optical Constants of Solids*, E. D. Palik, Ed., pp. 513–558. Academic Press, Burlington, 1997.

[14] V. M. Ustinov, A. E. Zhukov, A. Y. Egorov, and N. A. Maleev, Eds., *Quantum Dot Lasers*, Number 11 in Series on Semiconductor Science and Technology. Oxford University Press, Oxford, 2003.

[15] J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho, 'Quantum cascade laser', *Science*, **264**(5158):553–556, 1994.

[16] C. Sirtori, P. Kruck, S. Barbieri, P. Collot, J. Nagle, M. Beck, J. Faist, and U. Oesterle, '$GaAs/Al_xGa_{1-x}As$ quantum cascade lasers', *Appl. Phys. Lett.*, **73**(24):3486–3488, 1998.

[17] R. Köhler, A. Tredicucci, F. Beltram, H. E. Beere, E. H. Linfield, A. G. Davies, D. A. Ritchie, R. C. Iotti, and F. Rossi, 'Terahertz semiconductor-heterostructure laser', *Nature*, **417**:156–159, 2002.

[18] C. Sirtori, P. Kruck, S. Barbieri, H. Page, and J. Nagle, 'Low-loss Al-free waveguides for unipolar semiconductor lasers', *Appl. Phys. Lett.*, **75**(25):3911–3913, 1999.

[19] T. D. Visser, B. Demeulenaere, J. Haes, D. Lenstra, R. Baets, and H. Blok, 'Confinement and modal gain in dielectric waveguides', *J. Lightwave Technol.*, **14**:885, 1996.

[20] C. Sirtori, C. Gmachl, F. Capasso, J. Faist, D. L. Sivco, A. L. Hutchinson, and A. Y. Cho, 'Long-wavelength ($\lambda \approx 8$–$11.5\ \mu$m) semiconductor lasers with waveguides based on surface plasmons', *Opt. Lett.*, **23**(17):1366–1368, 1998.

[21] R. Colombelli, F. Capasso, C. Gmachl, A. L. Hutchinson, D. L. Sivco, A. Tredicucci, M. C. Wanke, A. M. Sergent, and A. Y. Cho, 'Far-infrared surface-plasmon quantum-cascade lasers at 21.5 μm and 24 μm wavelengths', *Appl. Phys. Lett.*, **78**(18):2620–2622, 2001.

[22] S. Adachi, Institution of Electrical Engineers, and INSPEC (Information Service), Eds., *Properties of Aluminium Gallium Arsenide*, Number 7 in EMIS Datareviews series. IEE, INSPEC, London, 1993.

[23] INSPEC (Information Service), Ed., *Properties of Indium Phosphide*, Number 6 in EMIS Datareviews Series. INSPEC, London; New York, 1991.

[24] A. Lops, V. Spagnolo, and G. Scamarcio, 'Thermal modeling of GaInAs/AlInAs quantum cascade lasers', *J. Appl. Phys.*, **100**(4):043109, 2006.

[25] C. Evans, D. Indjin, Z. Ikonic, P. Harrison, M. Vitiello, V. Spagnolo, and G. Scamarcio, 'Thermal modeling of terahertz quantum-cascade lasers: comparison of optical waveguides', *IEEE J. Quantum Electron.*, **44**(7):680–685, 2008.

[26] C. Kittel, *Introduction to Solid State Physics*, John Wiley & Sons, Inc., Hoboken, NJ, 8th edition, 2005.

[27] S. Adachi, *Physical Properties of III–V Semiconductor Compounds: InP, InAs, GaAs, GaP, InGaAs, and InGaAsP*, John Wiley & Sons, Inc., New York, 1992.

[28] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, Third edition, 1999.

[29] M. Salih, P. Dean, A. Valavanis, S. P. Khanna, L. H. Li, J. E. Cunningham, A. G. Davies, and E. H. Linfield, 'Terahertz quantum cascade lasers with thin resonant-phonon depopulation active regions and surface-plasmon waveguides', *J. Appl. Phys.*, **113**(11):113110, 2013.

[30] E. Strupiechonski, D. Grassani, D. Fowler, F. H. Julien, S. P. Khanna, L. Li, E. H. Linfield, A. G. Davies, A. B. Krysa, and R. Colombelli, 'Vertical subwavelength mode confinement in terahertz and mid-infrared quantum cascade lasers', *Appl. Phys. Lett.*, **98**(10):101101–101101–3, 2011.

[31] J. D. Hoffman, *Numerical Methods for Engineers and Scientists*, Marcel Dekker, New York, 2nd edition, 2001.

[32] M. S. Vitiello, G. Scamarcio, and V. Spagnolo, 'Time-resolved measurement of the local lattice temperature in terahertz quantum cascade lasers', *Appl. Phys. Lett.*, **92**(10):101116–101116–3, 2008.

[33] A. Valavanis, P. Dean, A. Scheuring, M. Salih, A. Stockhausen, S. Wuensch, K. Il'in, S. Chowdhury, S. P. Khanna, M. Siegel, A. G. Davies, and E. H. Linfield, 'Time-resolved measurement of pulse-to-pulse heating effects in a terahertz quantum cascade laser using an NbN superconducting detector', *Appl. Phys. Lett.*, **103**(6):061120–061120–4, 2013.

14

Multiband envelope function (k.p) method

Z. Ikonić

University of Leeds

14.1 Symmetry, basis states and band structure

The band structures of common semiconductors such as GaAs, InP, Si and Ge, which have the diamond/zinc blende crystal structure, are more complicated in the valence band than they are in the conduction band. This is related to the fact that, looking microscopically, at the level of the crystalline unit cell, the conduction-band states are mostly s-like in character (i.e. the microscopic 'Bloch' wave function has the symmetry of an atomic s orbital), while the valence-band states are mostly p-like in character. There are three degenerate p-type atomic orbitals, the symmetry of which is denoted as x, y and z, which stems from the direction along which the orbitals are aligned. It is therefore natural that all three of them will take part in valence-band state wave functions.

Detailed microscopic calculations (e.g. using the empirical pseudo-potential method; see Chapter 15) show that there are two valence bands degenerate at the centre of the Brillouin zone (the Γ-point), and close by (a few tens to a few hundreds of millielectronvolts below) there is a third band. The first two are called heavy-hole (HH) and light-hole (LH) bands, and they cease to be degenerate for non-zero values of the wave vector $\mathbf{k}$: the energy of the former descends at a slower rate as the wave vector moves away from the Γ-point, which corresponds to a larger effective mass, hence the name. The third band is called the spin–orbit split-off (SO) band. An example of the valence-band dispersion is given in Fig. 14.1.

Any one of the bands is an energy eigenstate of the bulk material, and 'pure' (single-band) states with definite energy may therefore exist in bulk. However, in the case of position-dependent potentials this will no longer be true: states in quantum wells, for instance, will be 'mixtures', with *all* the bulk bands contributing to their wave functions. The contribution of a particular bulk band to a quantised state generally depends on its energy spacing from

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

that band: the smaller the energy, the larger the relative contribution will be (this follows from quantum mechanical perturbation theory). When two or more bands are degenerate, or almost degenerate, states in their vicinity are likely to have similar contributions from these bands.

In a quantum-well-like structure, for energies which are not far from the conduction-band edges of the constituent materials, a quantised state wave function will have contributions mostly from the conduction bands of these materials, each of which has s-like character, and there will be a single envelope wave function (solution of the effective mass Schrödinger equation), which represents the amplitude of these s-like Bloch functions. On the other hand, a quantised state near the valence-band edges of any of the constituent materials is expected to comprise two or three of the bulk valence bands with comparable contributions, each having its own envelope function. Depending on the accuracy required in the calculation of the quantised states, and on the positions of the bands in the bulk materials, different numbers of bands may be included in the calculation. It may sometimes suffice to explicitly include just HH and LH bands, or, on other occasions, also the SO band in the description of the system.

Generally, the Hamiltonians that describe states in such situations are matrices, or systems of coupled Schrödinger equations, which will deliver the possible energies and wave functions expressed as a set of envelope functions (which vary slowly over a crystalline unit cell), themselves representing the amplitudes of the corresponding basis states (usually the bulk bands). For this reason, the method is known as the 'multiband envelope function', or 'multiband effective mass method' and, because the interaction of bulk bands is described via the **k.p** perturbation, it is also known as the **k.p** method.

Clearly, the concept of bulk band mixing in forming quantised states of a system applies to more remote bands as well. The conduction-band quantised states will thus include contributions from bulk valence-band states, and vice versa, and there exist extended versions of the **k.p** method that explicitly include the HH, LH, SO, and the conduction band, or still wider variants including even more remote bands. However, in this work attention will be focused on four- and six-band Hamiltonians, which explicitly include the valence-band states. The number of bands the Hamiltonian is named after is the number of bands that are explicitly included, i.e. their envelope wave functions are explicitly evaluated. However, such Hamiltonians do *implicitly* account for the existence of other, more remote bands, and their influence is incorporated via the values of the material parameters. In this context, the conventional (conduction-band) effective mass Schrödinger equation is just a special case of the multiband envelope function model, where the existence of bands other than the conduction band is accounted for by using the effective, rather than the free-electron mass, and only the conduction-band envelope wave function is calculated explicitly. The four- and six-band Hamiltonians for the valence envelope wave functions were derived by Luttinger and Kohn [1] using **k.p** perturbation theory, while the eight-band model (which includes the conduction band) was developed by Pidgeon and Brown [2].

14.2 Valence-band structure and the 6×6 Hamiltonian

It is beyond the scope of this book to derive the Hamiltonian that describes the valence-band states, so its form will be stated and it will be employed in the calculation of quantised

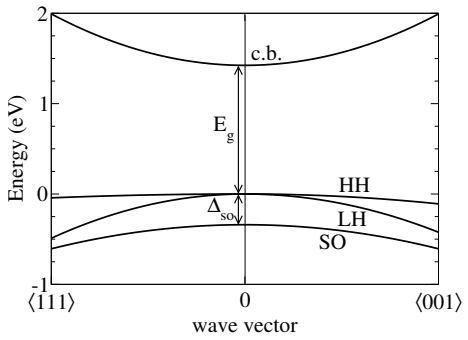


Figure 14.1: The valence and conduction bands of a group IV or III–V semiconductor near the Γ-point of the Brillouin zone

states within the valence band of nanostructures in order to illustrate how it is used. In many cases good accuracy can be obtained by using the so-called 6×6 Hamiltonian, although its shortened version, the 4×4 Hamiltonian is frequently just as good—it all depends on the energy range of interest. As mentioned above, a valid basis set can be the atomic p-like orbitals with x-, y-, and z-like spatial symmetry, here denoted as $|X\rangle$, $|Y\rangle$, $|Z\rangle$, each also having the spin projection along the z-axis equal to either $+1/2$ or $-1/2$, denoted as $\uparrow$ or $\downarrow$, respectively. However, common practice is to change from the $|X\uparrow\rangle$, $|Y\uparrow\rangle$, $|Z\uparrow\rangle$, $|X\downarrow\rangle$, $|Y\downarrow\rangle$, $|Z\downarrow\rangle$ basis into another one, such that its member functions are simultaneously the eigenstates of the angular momentum operator (with eigenvalues J equal to $3\hbar/2$ or $\hbar/2$), and of its projection along the z-axis m_J (with eigenvalues equal to $\pm 3/2$ or $\pm 1/2$). This is achieved by making appropriate linear combinations of the atomic basis states. The list of this new set of $|J, m_J\rangle$ basis states is given below in equation (14.1).

The precise form of the Hamiltonian depends on the prefactors in equation (14.1), for example, the presence of the imaginary number i or -1 does not change the state properties, and it also depends on how these states are ordered in the list. There is no unique choice that is universally accepted in the literature; however, one of the frequently used possibilities (see, for example, [3]), which will be adopted here, reads:

$$
\begin{aligned}
|3/2, 3/2\rangle &= (1/\sqrt{2})\,|(X + iY)\uparrow\rangle \\
|3/2, -3/2\rangle &= (1/\sqrt{2})\,|(X - iY)\downarrow\rangle \\
|3/2, 1/2\rangle &= (1/\sqrt{6})\,|(X + iY)\downarrow\rangle - (\sqrt{2}/\sqrt{3})|Z\uparrow\rangle \\
|3/2, -1/2\rangle &= -(1/\sqrt{6})\,|(X - iY)\uparrow\rangle - (\sqrt{2}/\sqrt{3})\,|Z\downarrow\rangle \\
|1/2, 1/2\rangle &= (1/\sqrt{3})\,|(X + iY)\downarrow\rangle + (1/\sqrt{3})|Z\uparrow\rangle \\
|1/2, -1/2\rangle &= -(1/\sqrt{3})\,|(X - iY)\uparrow\rangle + (1/\sqrt{3})\,|Z\downarrow\rangle
\end{aligned}
\tag{14.1}
$$

Just as with the original $(|X\uparrow\rangle, |Y\uparrow\rangle, |Z\uparrow\rangle, |X\downarrow\rangle, |Y\downarrow\rangle, |Z\downarrow\rangle)$ states, the new basis states are all orthogonal to each other. In this $|J, m_J\rangle$ basis the 6×6 Hamiltonian that

describes the HH, LH and SO bands for the bulk reads:

$$\mathsf{H} = \begin{pmatrix} P+Q & 0 & -S & R & (1/\sqrt{2})S & \sqrt{2}R \\ 0 & P+Q & -R^\dagger & -S^\dagger & -\sqrt{2}R^\dagger & (1/\sqrt{2})S^\dagger \\ -S^\dagger & -R & P-Q & 0 & \sqrt{2}Q & \sqrt{3/2}S \\ R^\dagger & -S & 0 & P-Q & -\sqrt{3/2}S^\dagger & \sqrt{2}Q \\ (1/\sqrt{2})S^\dagger & -\sqrt{2}R & \sqrt{2}Q & -\sqrt{3/2}S & P+\Delta_{\mathrm{SO}} & 0 \\ \sqrt{2}R^\dagger & (1/\sqrt{2})S & \sqrt{3/2}S^\dagger & \sqrt{2}Q & 0 & P+\Delta_{\mathrm{SO}} \end{pmatrix} \tag{14.2}$$

where

$$\begin{aligned} P &= \frac{\hbar^2}{2m_0}\gamma_1(k_x^2+k_y^2+k_z^2) \\ Q &= \frac{\hbar^2}{2m_0}\gamma_2(k_x^2+k_y^2-2k_z^2) \\ R &= \frac{\hbar^2}{2m_0}\sqrt{3}\left[-\gamma_2(k_x^2-k_y^2)+2i\gamma_3 k_x k_y\right] \\ S &= \frac{\hbar^2}{2m_0}2\sqrt{3}\gamma_3 k_z k_- \end{aligned} \tag{14.3}$$

where $k_- = k_x - ik_y$, $\gamma_{1,2,3}$ are the Luttinger parameters, and Δ_{SO} is the spin–orbit splitting—the spacing between the HH (or LH) band and the SO band at the ($k = 0$) Γ-point. The values of $\gamma_{1,2,3}$ and Δ_{SO} in some common semiconductors are given in Table 14.1.

In writing equation (14.2) the convention has been used that the hole energy is measured from the top of the valence band downwards (the inverted energy picture). This is because it is usually easier to look at the hole band structure in the same manner as that for electrons, and this is possible if only holes are considered. For the true energy picture (as in Fig. 14.1), all the terms in the Hamiltonian should be multiplied by -1. Furthermore, it is important to note that the coordinate system in which this Hamiltonian is written is not orientated arbitrarily—the axes x, y and z are aligned along the edges of the crystalline cubic unit cell.

The Schrödinger equation corresponding to this Hamiltonian may still be written as

$$\mathcal{H}\psi = E\psi \tag{14.4}$$

but $\psi = \psi(\mathbf{r})$ is a *vector* (a six-component array). Note that the Hamiltonian is a Hermitian matrix with simple scalars as its elements (which depend on the material parameters) and wave-vector components k_x, k_y and k_z. If the eigenenergies and the corresponding eigenvectors are found, the latter will obviously be lists of (possibly complex-valued) scalar constants. What would these mean? It is implicitly assumed that the wave function has a plane wave form, i.e. all the six components of ψ have the common, plane wave type of spatial behaviour, i.e. $\sim \exp(i\mathbf{k}.\mathbf{r})$. This makes them the components of the envelope wave function and, if interested in a more detailed form, each of them multiplies a corresponding basis state from the list in equation (14.1) and can be added together to construct the 'true' microscopic wave function of a state with energy E. The eigenvectors are generally 'full', i.e. their entries usually have non-zero values, which means that a plane wave state has all the

Table 14.1 Parameters relevant for the valence-band structure in a few semiconductors [4, 5]

	Si	Ge	GaAs	AlAs	InAs
γ_1	4.22	13.4	6.98	3.76	20.0
γ_2	0.39	4.25	2.06	0.82	8.5
γ_3	1.44	5.69	2.93	1.42	9.2
Δ_{SO} (eV)	0.044	0.290	0.341	0.28	0.39
a_v (eV)	2.46	1.24	-1.116	-2.47	-1.00
b (eV)	-2.10	-2.86	-2.0	-2.3	-1.8

six basis states admixed. This is in contrast to the conduction band, where one usually deals with 'pure' spin-up or spin-down states.

To have the usual meaning of a wave function, an eigenvector ψ has to be normalised to unity, i.e. all the components of a vector may have to be multiplied by a suitable constant so that:

$$\psi^{\dagger}\psi = 1 \tag{14.5}$$

is satisfied. The † symbol denotes the Hermitian conjugate, i.e. the transpose (a column vector becomes a row vector) followed by the complex conjugate (take the complex conjugate of all the elements).

In order to find eigenenergies for a specified **k**, solutions can be sought to $\det|\mathsf{H} - E\mathsf{I}| = 0$, which delivers a sixth-order polynomial in the energy E. This could only be solved numerically, unless it is noted that it can be factored into two identical third-order polynomials, thus allowing for analytic, though lengthy, solutions. However, the situation is quite simple at the zone centre, i.e. if we set $k_x = k_y = k_z = 0$ the Hamiltonian becomes:

$$\mathsf{H}(\mathbf{k}=0) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \Delta_{\mathrm{SO}} & 0 \\ 0 & 0 & 0 & 0 & 0 & \Delta_{\mathrm{SO}} \end{pmatrix} \tag{14.6}$$

This has a fourfold degenerate eigenvalue $E = 0$, and a twofold degenerate eigenvalue $E = \Delta_{\mathrm{SO}}$, consistent with the concept that there are HH and LH branches degenerate at $\mathbf{k} = 0$, and an SO branch displaced by Δ_{SO} from them. As for the eigenvectors: for each of the four degenerate values $E = 0$, just one of the four amplitudes of $|3/2, 3/2\rangle$, $|3/2, -3/2\rangle$, $|3/2, 1/2\rangle$, $|3/2, -1/2\rangle$ has value unity, the others being equal to zero (though, due to degeneracy, any other linear combination of these states would do just as well). Similarly, the eigenvectors corresponding to the two degenerate eigenvalues $E = \Delta_{\mathrm{SO}}$ may be taken to have non-zero amplitude of the $|1/2, 1/2\rangle$, and of the $|1/2, -1/2\rangle$ states. For finite values of **k**, the fourfold degeneracy splits into two pairs of doubly degenerate states (HH and LH have different energies), so all three branches remain doubly degenerate—in analogy to the spin degeneracy of conduction-band states.

14.3 4×4 valence-band Hamiltonian

For non-zero **k**, the expressions for the eigenenergies would be quite lengthy, so consider in more detail a 'shortened' version, i.e. a 4×4 Hamiltonian that includes only the HH and LH states in its basis. This is obtained by excluding the $|1/2, \pm 1/2\rangle$ states, i.e. by removing the fifth and sixth rows and columns from the Hamiltonian matrix in equation (14.2):

$$\mathsf{H} = \begin{pmatrix} P+Q & 0 & -S & R \\ 0 & P+Q & -R^\dagger & -S^\dagger \\ -S^\dagger & -R & P-Q & 0 \\ R^\dagger & -S & 0 & P-Q \end{pmatrix} \tag{14.7}$$

Finding $\det|\mathsf{H} - E\mathsf{I}|$ gives a fourth-order polynomial in E, but it is easy to see that it is the square of a quadratic polynomial:

$$(P+Q-E)(P-Q-E) - |R|^2 - |S|^2 = 0 \tag{14.8}$$

Using the expressions in equation (14.3), it can be found that the two solutions read:

$$E_\pm(k_x, k_y, k_z) = \frac{\hbar^2}{2m_0}\left[\gamma_1 k^2 \pm 2\sqrt{\gamma_2^2 k^4 - 3(\gamma_2^2 - \gamma_3^2)(k_x^2 k_y^2 + k_x^2 k_z^2 + k_y^2 k_z^2)}\right] \tag{14.9}$$

where $k^2 = k_x^2 + k_y^2 + k_z^2$.

It is immediately clear from equation (14.9) that both solutions, also called dispersion branches, are non-parabolic (E does not depend only on k^2) and anisotropic (E depends on $k_{x,y,z}$ in a manner different from just that of its modulus k, i.e. the energy varies differently in different directions of the wave-vector space). This latter property is also called 'band warping'.

Consider now the special case when two components of the wave vector are zero, e.g. $k_x = k_y = 0$, while $k_z \neq 0$. From the definitions in equation (14.3), $R = S = 0$, hence:

$$E_+ = k_z^2\hbar^2(\gamma_1 + 2\gamma_2)/2m_0 \quad \text{and} \quad E_- = k_z^2\hbar^2(\gamma_1 - 2\gamma_2)/2m_0 \tag{14.10}$$

Clearly, the E_- branch is characterised by an effective mass in the z-direction equal to $m_0/(\gamma_1 - 2\gamma_2)$, and this is larger than the effective mass of the E_+ branch, which amounts to $m_0/(\gamma_1 + 2\gamma_2)$. For this reason, the E_- branch is called the 'heavy hole' (HH) branch, and its energy is denoted as E_{HH}, while the E_+ branch is referred to as the 'light hole' (LH) branch, and its energy denoted as E_{LH}. Under these conditions, where in fact $E_{\text{HH}} = P + Q$, the first two equations in $(\mathcal{H} - E_{\text{HH}})\psi = 0$ have all zero entries, while the third and fourth equations do not. From this it can be concluded that elements 3 and 4 in the vector ψ are zero, while elements 1 and 2 are not. Therefore, $[1\ \ 0\ \ 0\ \ 0]^T$ can be chosen for one HH state vector, and $[0\ \ 1\ \ 0\ \ 0]^T$ for the other (remember that E_{HH} is doubly degenerate), though any pair of linear combinations of these two vectors would do just as well. However, if choosing to make such linear combinations, care should be taken to make them mutually orthogonal in order to avoid complications in any further use of the wave functions. In any case, for $k_z \neq 0$, the HH branch has only $|3/2, \pm 3/2\rangle$ basis states as its constituents. Similar reasoning for the E_{LH} branch shows that, for $k_z \neq 0$, the LH branch includes only $|3/2, \pm 1/2\rangle$ states.

For non-zero k_x and/or k_y, however, the system of linear equations is generally full, which indicates that a hole state has a finite contribution from all the basis states. This may seem puzzling, in view of the fact that there is essentially no difference between the x- and z-directions in the bulk crystal. However, this difference arises because of the choice of basis states: each of them has a definite projection of momentum and spin along the z-axis, and therefore (as is known from quantum mechanics) does not have the same property along the x-axis.

14.4 Complex band structure

While equation (14.8) and its solutions in equation (14.9) give the possible values of hole energy for a particular wave vector **k**, it is interesting to consider the reverse problem: what values of k_z may holes have if their energy E and the other two components of the wave vector (k_x and k_y) are specified? For the sake of simplicity, k_y is set to zero, while k_x and E are generally non-zero and are real. Equation (14.8) in its expanded form then reads:

$$\left(\frac{\hbar^2 k_z^2}{2m_0}\right)^2 (\gamma_1^2 - \gamma_2^2) + \left[(2\gamma_1^2 + 8\gamma_2^2 - 12\gamma_3^2)\frac{\hbar^2 k_x^2}{2m_0} - 2E\gamma_1\right]\frac{\hbar^2 k_z^2}{2m_0} + \left[E - (\gamma_1 + 2\gamma_2)\frac{\hbar^2 k_x^2}{2m_0}\right]\left[E - (\gamma_1 - 2\gamma_2)\frac{\hbar^2 k_x^2}{2m_0}\right] \equiv Ak_z^4 + Bk_z^2 + C = 0 \quad (14.11)$$

which is, of course, quadratic in k_z^2, with the two solutions:

$$(k_z^2)_{1,2} = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A} \quad (14.12)$$

each of which delivers two k_zs as a positive and a negative root of k_z^2. In real semiconductors the values of the Luttinger parameters γ_1 and γ_2 are such that A is always positive, while B and C may be of either sign, depending on E and k_x, as well as the material parameters.

Consider the case of $4AC < 0$, i.e. $C < 0$. From the form used to write C, it is clear that this will happen when k_x and E are such that $(\gamma_1 - \gamma_2)k_x^2 < 2m_0E/\hbar^2 < (\gamma_1 + 2\gamma_2)k_x^2$ is satisfied. The $\sqrt{B^2 - 4AC}$ is then larger than B, and regardless of the sign of B one of the roots $(k_z^2)_{1,2}$ will be positive and the other one negative. Hence there will be a pair of real and a pair of purely imaginary k_z values.

Consider now the case of $4AC > B^2$: equation (14.12) then contains the square root of a negative number, which will imply that both roots will be fully complex numbers (with both the real and imaginary parts non-zero). Finally, for $0 < 4AC < B^2$, both values of $(k_z^2)_{1,2}$ are either positive or negative, depending on the sign of B, i.e. all roots are either real or imaginary.

Real-valued solutions for k_z imply conventional plane wave envelope wave functions, which is an allowed state in an infinite bulk. Complex k_z implies an 'evanescent' wave, which decays in one direction and increases in the opposite direction, and which may simultaneously oscillate (if the real part of $k_z \neq 0$). Because of the infinite length in the z-direction, and the inability to normalise the wave function, such states are not allowed in an infinite bulk crystal, but are perfectly allowed in finite regions, just as is the case with electrons, and can therefore appear in structures of finite extent.

The above considerations were presented for the 4×4 model; however, the general conclusions about the possible types of evanescent waves remain for more elaborate models as well. The list of all possible states at a particular energy E that behave exponentially is called the 'complex band structure'. The wave functions vary along a particular direction, say z, as $\exp(ik_z z)$, where k_z is a real or complex wave vector.

The case of electrons in the conduction band is quite different as they have a scalar effective mass and really a very simple complex band structure: their wave vector can be only real or imaginary, $k_z = \pm\sqrt{2m^*E/\hbar^2 - k_x^2 - k_y^2}$, but never fully complex.

For holes, this situation only occurs at $k_x = k_y = 0$, when $\hbar^2(k_z^2)_1/2m_0 = E/(\gamma_1 - 2\gamma_2)$ and $\hbar^2(k_z^2)_2/2m_0 = E/(\gamma_1 + 2\gamma_2)$, so all k_zs are real for $E > 0$ and imaginary for $E < 0$.

14.5 Block-diagonalisation of the Hamiltonian

There is an elegant method, which allows one to simplify the 6×6 Hamiltonian, by recasting it into a block-diagonal form: it then has two 3×3 matrices as its diagonal elements and the remaining two off-diagonal elements are zero matrices. This is achieved by a suitable change of basis, i.e. by creating a new basis from linear combinations of the existing basis states. Essentially, this means that there are two completely independent sets of states, which are eigenstates of either one or the other 3×3 block. Since this block-diagonalisation reduces the size of the system to be considered at any time, it is useful when attempting to obtain a result in analytical form. Block-diagonalisation can also be useful in cases that are more complicated than straightforward bulk material, for example, strained bulk or two-dimensional quantum well systems.

The derivation of the transformation of the Hamiltonian into a block-diagonal form will not be reproduced here, the result will merely be stated, i.e. upon introducing the new basis:

$$\begin{aligned}
|F_1\rangle &= \alpha\ |3/2, -3/2\rangle - \alpha^*\ |3/2, 3/2\rangle \\
|F_2\rangle &= \beta\ |3/2, 1/2\rangle + \beta^*\ |3/2, -1/2\rangle \\
|F_3\rangle &= \beta\ |1/2, 1/2\rangle + \beta^*\ |1/2, -1/2\rangle \\
|F_4\rangle &= \alpha\ |3/2, -3/2\rangle + \alpha^*\ |3/2, 3/2\rangle \\
|F_5\rangle &= \beta\ |3/2, 1/2\rangle - \beta^*\ |3/2, -1/2\rangle \\
|F_6\rangle &= \beta\ |1/2, 1/2\rangle - \beta^*\ |1/2, -1/2\rangle
\end{aligned} \tag{14.13}$$

where:

$$\begin{aligned}
\alpha = \frac{1}{\sqrt{2}}\exp[i(\phi/2 + \eta + \pi/4)], \qquad \beta = \frac{1}{\sqrt{2}}\exp[i(\phi/2 - \eta - 3\pi/4)] \\
\phi = \arctan(k_y/k_x), \qquad \eta = \frac{1}{2}\arctan[(\gamma_3/\gamma_2)\tan(2\phi)]
\end{aligned} \tag{14.14}$$

the Hamiltonian takes a block-diagonal form:

$$\mathsf{H} = \begin{pmatrix} \mathsf{H}_+ & 0 \\ 0 & \mathsf{H}_- \end{pmatrix} \tag{14.15}$$

The two 3×3 blocks read:

$$\mathsf{H}_{\pm} = \begin{pmatrix} P+Q & R \mp iS & \sqrt{2}R \pm iS/\sqrt{2} \\ R \pm iS^{\dagger} & P-Q & \sqrt{2}Q \mp i\sqrt{3/2}S \\ \sqrt{2}R \mp iS^{\dagger}/\sqrt{2} & \sqrt{2}Q \pm i\sqrt{3/2}S^{\dagger} & P+\Delta_{\mathrm{SO}} \end{pmatrix} \tag{14.16}$$

where P and Q are the same as above, while R and S now read:

$$R = -\sqrt{3}\frac{\hbar^2}{2m_0}\gamma_{\phi}k_{||}^2, \qquad S = 2\sqrt{3}\frac{\hbar^2}{2m_0}\gamma_3 k_{||}k_z$$
$$\gamma_{\phi} = \sqrt{\overline{\gamma}^2 + \mu^2 - 2\overline{\gamma}\mu\cos\phi}, \qquad \overline{\gamma} = \frac{1}{2}(\gamma_2+\gamma_3), \qquad \mu = \frac{1}{2}(\gamma_3-\gamma_2) \tag{14.17}$$

where $k_{||}^2 = k_x^2 + k_y^2$ is the in-plane wave vector.

Without going into detail, it should be noted that, in order to make this transformation possible, the element R from equation (14.2) had to be approximated as given in equation (14.17). A simpler approximation, with $\gamma_{\phi} = (\gamma_2+\gamma_3)/2$, has previously been in use, but the one given above is much better. The important point to note is that the transformation in equation (14.13) does not depend on k_z, and when this becomes an operator $k_z \to -i\partial/\partial z$ (for application to quantum wells) there are no difficulties in changing from one basis to another.

To find the band energies and wave functions, the block-diagonal form of the Hamiltonian may then be used. Solving one 3×3 block at a time, the wave function is now a vector of length 3, i.e.

$$\mathsf{H}_{\pm}\psi_{\pm} = E\psi_{\pm}$$
$$\psi_{+}(z) = \begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix}, \qquad \psi_{-}(z) = \begin{pmatrix} F_4 \\ F_5 \\ F_6 \end{pmatrix} \tag{14.18}$$

Block-diagonalisation may also be applied to the 4×4 Hamiltonian of equation (14.7), using the basis states F_1, F_2, F_4, and F_5 defined in equation (14.13), which transforms the full 4×4 matrix into one which has two 2×2 blocks. Alternatively, one can simply cut out the third row and third column from the two blocks defined in equation (14.16) to obtain the same result. Taking, for example, the upper 2×2 block:

$$\begin{pmatrix} P+Q-E & R-iS \\ R+iS & P-Q-E \end{pmatrix}\begin{pmatrix} F_1 \\ F_2 \end{pmatrix} = \mathbf{0} \tag{14.19}$$

and evaluating $\det|\mathsf{H}_{+} - E\mathsf{I}| = 0$ gives:

$$E^2 - 2PE + P^2 - Q^2 - |R-iS|^2 = 0 \tag{14.20}$$

which has two non-degenerate solutions:

$$E_{\pm} = P \pm \sqrt{Q^2 + |R-iS|^2} \tag{14.21}$$

in which it can immediately be recognised that $E_{-} = E_{\mathrm{HH}}$ and $E_{+} = E_{\mathrm{LH}}$ from the previous section. The same values would be obtained from the lower 2×2 block.

Putting these solutions (one at a time!) back into either of the two equations of the homogeneous linear system in equation (14.19), as convenient, and solving it, gives the eigenvectors for the two eigenenergies as:

$$\begin{pmatrix} F_1 \\ F_2 \end{pmatrix}_{E_-} = \begin{pmatrix} P - Q - E_- \\ -(R + iS) \end{pmatrix} \qquad \begin{pmatrix} F_1 \\ F_2 \end{pmatrix}_{E_+} = \begin{pmatrix} R - iS \\ E_+ - P - Q \end{pmatrix} \tag{14.22}$$

These eigenvectors are not yet normalised to unity, which can be done simply by dividing each by the square root of the sum of its component moduli squared. In the special case when $k_x = k_y = 0$ and $k_z \neq 0$, the HH state vector is simply $[1 \quad 0]^T$ (i.e. only a particular linear combination of $|3/2, \pm 3/2\rangle$ states is present), while the LH state vector is $[0 \quad 1]^T$ (only $|3/2, \pm 1/2\rangle$ states are present).

14.6 The valence band in strained cubic semiconductors

In Chapter 7 it was shown that the application of mechanical strain leads to changes in the energy band structure in a crystal, although the discussion was limited to conduction-band states. In this section, the influence of strain on the valence-band structure will be considered.

As in Chapter 7, the strain may be described by six components, $\epsilon_1, \ldots, \epsilon_6$, corresponding to compression or expansion along each of the crystal axes ($\epsilon_1, \ldots, \epsilon_3$) and shear strain components ($\epsilon_4, \ldots, \epsilon_6$). The various terms in the Hamiltonian (see equation (14.2)) then acquire strain-dependent contributions (see, for example, [6]):

$$\begin{aligned} P \to P + P_\epsilon\,, \qquad & P_\epsilon = -a_v(\epsilon_1 + \epsilon_2 + \epsilon_3) \\ Q \to Q + Q_\epsilon\,, \qquad & Q_\epsilon = -\frac{b}{2}(\epsilon_1 + \epsilon_2 - 2\epsilon_3) \\ R \to R + R_\epsilon\,, \qquad & R_\epsilon = \frac{\sqrt{3}}{2} b(\epsilon_1 - \epsilon_2) - id\epsilon_6 \\ S \to S + S_\epsilon\,, \qquad & S_\epsilon = -d(\epsilon_5 - i\epsilon_4) \end{aligned} \tag{14.23}$$

where a_v, b and d are the Pikus–Bir deformation potentials, describing the influence of hydrostatic, uniaxial and shear strain, respectively.

The (in practice most important) case of biaxial strain generated by lattice-mismatched growth of semiconductor layers on a $[001]$-oriented substrate was considered in Section 7.2. In this case, the shear-strain components $\epsilon_4, \epsilon_5, \epsilon_6 = 0$ and hence the shear deformation potential d in equation (14.23) is not required. The in-plane strain components $\epsilon_1 = \epsilon_2$ are given by equation (7.18) and the perpendicular strain by equation (7.21).

Looking again at the special case of $\mathbf{k} = 0$, the Hamiltonian takes the form:

$$\mathsf{H} = \begin{pmatrix} P_\epsilon + Q_\epsilon & 0 & 0 & 0 & 0 & 0 \\ 0 & P_\epsilon + Q_\epsilon & 0 & 0 & 0 & 0 \\ 0 & 0 & P_\epsilon - Q_\epsilon & 0 & \sqrt{2}Q_\epsilon & 0 \\ 0 & 0 & 0 & P_\epsilon - Q_\epsilon & 0 & \sqrt{2}Q_\epsilon \\ 0 & 0 & \sqrt{2}Q_\epsilon & 0 & P_\epsilon + \Delta_{\mathrm{SO}} & 0 \\ 0 & 0 & 0 & \sqrt{2}Q_\epsilon & 0 & P_\epsilon + \Delta_{\mathrm{SO}} \end{pmatrix} \tag{14.24}$$

This is a bit more complicated than the unstrained case. For example, the presence of non-zero off-diagonal matrix elements on positions that link LH and SO states implies that these two bands are coupled even at $\mathbf{k} = 0$; however, the HH band remains independent. Away from the zone centre, however, all the three bands are coupled. If the size of the system is reduced to a 4×4 Hamiltonian, by removing the $|1/2, \pm 1/2\rangle$ states (the fifth and sixth rows and columns), a simple diagonal matrix is obtained and its eigenenergies can readily be written, each of which is twofold degenerate:

$$E_{\text{HH}}(0) = P_\epsilon + Q_\epsilon, \quad E_{\text{LH}}(0) = P_\epsilon - Q_\epsilon \tag{14.25}$$

Consider the case of compressive in-plane strain in a layer (i.e. $\epsilon_1 < 0$). Therefore, according to equation (7.21), the result of this accommodation of the in-plane lattice constant to the value of the substrate is that ϵ_3 must be greater than zero. The result is a finite value of P_ϵ which, as it only occurs along the diagonal terms in equations (14.2) or (14.15), has the trivial effect in bulk of a rigid energy shift of all the bands by that amount, with no other physical consequences.

The term Q_ϵ, however, has non-trivial effects. The value of b in semiconductors is negative, therefore $Q_\epsilon < 0$, and the HH band edge will decrease, while the LH band edge will increase (with the inverted energy picture), hence the HH and LH bands cease to be degenerate, with the light holes at the zone centre having an energy greater than that of the heavy holes. In the opposite case of tensile in-plane strain, the above effects are reversed.

Returning to the 6×6 Hamiltonian, the eigenvalue equation is a sixth-order polynomial in the energy E. It can be written as the square of a third-order polynomial, which can be solved analytically, but this is not an easy thing to do. Again, looking at the special case of $\mathbf{k} = 0$ summarised in equation (14.24), notice that the diagonal elements $P_\epsilon + Q_\epsilon$ stand alone in both their rows and columns, which means that this is a doubly degenerate eigenvalue and can be factored out. The remaining fourth-order polynomial problem is then easier to solve and, in fact, it can be seen that it is the square of a quadratic polynomial. The eigenenergies, which are the edges of the three valence bands in a strained semiconductor, follow as:

$$\begin{aligned}
E_{\text{HH}}(0) &= P_\epsilon + Q_\epsilon \\
E_{\text{LH}}(0) &= P_\epsilon - \frac{1}{2}\left(Q_\epsilon - \Delta_{\text{SO}} + \sqrt{\Delta_{\text{SO}}^2 + 2\Delta_{\text{SO}}Q_\epsilon + 9Q_\epsilon^2}\right) \\
E_{\text{SO}}(0) &= P_\epsilon - \frac{1}{2}\left(Q_\epsilon - \Delta_{\text{SO}} - \sqrt{\Delta_{\text{SO}}^2 + 2\Delta_{\text{SO}}Q_\epsilon + 9Q_\epsilon^2}\right)
\end{aligned} \tag{14.26}$$

It is important to note, for the purpose of a later discussion on the band offset at an interface between different materials, that the average ('centre of mass') of these three valence-band edges shifts only because of the hydrostatic component P_ϵ, from its unstrained value of $\Delta_{\text{SO}}/3$. In the case of a large spin–orbit splitting ($\Delta_{\text{SO}} \gg Q_\epsilon$), by taking the first two terms in Taylor series expansions of the roots in equation (14.26), the results in equation (14.25) are recovered, while the SO band in this limit appears relatively insensitive to strain.

Values of the parameters necessary for strain calculations in a few common semiconductors are given in Table 14.1.

14.7 Hole subbands in heterostructures

In a heterostructure, where the material composition is modulated, and hence all the material parameters, as well as the valence-band edge energies, become position-dependent, the Hamiltonian in equation (14.2) has to be modified appropriately. For some time it has been accepted that equation (14.2) is modified in analogy with the effective mass Schrödinger equation: in a two-dimensional system, for example, k_x and k_y remain as they are, because only a plane wave type of solution is allowed in the non-quantising[1] directions, along which the potential is constant; however, the wave vector k_z along the quantising (growth) direction is substituted for its quantum mechanical operator $-i\partial/\partial z$, and then, in order to preserve the hermiticity of the Hamiltonian, the terms of type γk_z^2 become $k_z\gamma k_z$, while the terms of type γk_z become $(\gamma k_z + k_z\gamma)/2$ (note that the γs are position-dependent in a heterostructure).

Furthermore, the diagonal elements of the Hamiltonian are amended with the potential V, which could have contributions from the valence-band offset in the particular material, the potential from an external electrostatic field, or the self-consistent space-charge electrostatic potential. Following this, it was shown by Foreman [3] that further modifications are necessary, which improve the accuracy of the method, and under these developments the Hamiltonian now reads:

$$\mathsf{H} = \begin{pmatrix} P+Q+V & 0 & -S_- & R & (1/\sqrt{2})S_- & \sqrt{2}R \\ 0 & P+Q+V & -R^\dagger & -S_+ & -\sqrt{2}R^\dagger & (1/\sqrt{2})S_+ \\ -S_-^\dagger & -R & P-Q+V & C & \sqrt{2}Q & \sqrt{3/2}\Sigma_- \\ R^\dagger & -S_+^\dagger & C^\dagger & P-Q+V & -\sqrt{3/2}\Sigma_+ & \sqrt{2}Q \\ (1/\sqrt{2})S_-^\dagger & -\sqrt{2}R & \sqrt{2}Q & -\sqrt{3/2}\Sigma_+^\dagger & P+\Delta_{\mathrm{SO}}+V & -C \\ \sqrt{2}R^\dagger & (1/\sqrt{2})S_+^\dagger & \sqrt{3/2}\Sigma_-^\dagger & \sqrt{2}Q & -C^\dagger & P+\Delta_{\mathrm{SO}}+V \end{pmatrix} \tag{14.27}$$

where:

$$P = \frac{\hbar^2}{2m_0}\gamma_1(k_x^2+k_y^2+k_z^2)$$

$$Q = \frac{\hbar^2}{2m_0}\gamma_2(k_x^2+k_y^2-2k_z^2)$$

$$R = \sqrt{3}\frac{\hbar^2}{2m_0}(-\overline{\gamma}k_-^2+\mu k_+^2)$$

$$S_\pm = 2\sqrt{3}\frac{\hbar^2}{2m_0}k_\pm\left[(\sigma-\delta)k_z + k_z\pi\right]$$

$$\Sigma_\pm = 2\sqrt{3}\frac{\hbar^2}{2m_0}k_\pm\left\{\left[\frac{1}{3}(\sigma-\delta)+\frac{2}{3}\pi\right]k_z + k_z\left[\frac{2}{3}(\sigma-\delta)+\frac{1}{3}\pi\right]\right\}$$

$$C = 2\frac{\hbar^2}{2m_0}k_-\left[k_z(\sigma-\delta-\pi)-(\sigma-\delta-\pi)k_z\right]$$

$$k_\pm = k_x \pm ik_y, \qquad \overline{\gamma} = \frac{1}{2}(\gamma_2+\gamma_3), \qquad \mu = \frac{1}{2}(\gamma_3-\gamma_2)$$

[1] In the plane of the quantum well.

$$\sigma = \overline{\gamma} - \frac{1}{2}\delta, \qquad \pi = \mu + \frac{3}{2}\delta, \qquad \delta = \frac{1}{9}(1 + \gamma_1 + \gamma_2 - 3\gamma_3) \tag{14.28}$$

It is possible to show that equation (14.27) reduces to equation (14.2) if the Luttinger parameters (γ_1, γ_2 and γ_3) are set as constants.

As for the case of the bulk Hamiltonian, equation (14.27) can also be block-diagonalised into two 3×3 blocks, which read:

$$\mathsf{H}_{\pm} = \begin{pmatrix} P+Q+V & R \mp iS & \sqrt{2}R \pm iS/\sqrt{2} \\ R \pm iS^{\dagger} & P-Q \mp iC+V & \sqrt{2}Q \mp i\sqrt{3/2}\Sigma \\ \sqrt{2}R \mp iS^{\dagger}/\sqrt{2} & \sqrt{2}Q \pm i\sqrt{3/2}\Sigma^{\dagger} & P+\Delta_{\mathrm{SO}} \pm iC+V \end{pmatrix} \tag{14.29}$$

where P and Q are the same as above, but R, S, Σ and C now read:

$$R = -\sqrt{3}\frac{\hbar^2}{2m_0}\gamma_\phi k_{||}^2, \qquad S = 2\sqrt{3}\frac{\hbar^2}{2m_0}k_{||}\left[(\sigma-\delta)k_z + k_z\pi\right]$$

$$\Sigma = 2\sqrt{3}\frac{\hbar^2}{2m_0}k_{||}\left\{\left[\frac{1}{3}(\sigma-\delta)+\frac{2}{3}\pi\right]k_z + k_z\left[\frac{2}{3}(\sigma-\delta)+\frac{1}{3}\pi\right]\right\}$$

$$C = 2\frac{\hbar^2}{2m_0}k_{||}\left[k_z(\sigma-\delta-\pi) - (\sigma-\delta-\pi)k_z\right]$$

$$\gamma_\phi = \sqrt{\overline{\gamma}^2 + \mu^2 - 2\overline{\gamma}\mu\cos 4\phi}, \qquad k_{||}^2 = k_x^2 + k_y^2 \tag{14.30}$$

It would now appear unclear what values of γ_2 and γ_3 should be used in γ_ϕ; however, it is a good approximation to use their average values across the structure.

For a layer-type structure (i.e. with a constant composition inside any of the layers) use equation (14.27) or (14.29), to find the boundary conditions for the wave function at interfaces. This is achieved by formal integration across the interface, in the same way as with the effective mass Schrödinger equation (resulting in the conclusion that ψ, as well as $(1/m^*)\partial\psi/\partial z$, is conserved across the interface). Such integration of equation (14.27) or (14.29) shows that the amplitudes F_i in the wave-function vector are individually conserved, and also that there are particular linear combinations of both the derivatives and the amplitudes of all F_i components that are conserved across the interface (see later).

Methods of calculating the eigenstates of heterostructures may be divided into two groups. One, which is practical only for one-dimensional heterostructure potentials (like quantum wells), uses a 'layer approach' and first finds the relevant properties of each single layer in the structure, before proceeding to find its 'global' properties—in particular, the eigenstates.

The other approach considers directly the heterostructure as a whole. It is computationally more demanding than the layer approach, but is equally applicable to quantum wells, wires and dots. In either case, the final ingredient we need for the calculation is the position-dependent potential $V(z)$ (or $V(\mathbf{r})$ in multidimensional structures) that is to be used in the Hamiltonian equation (14.27) or (14.29).

14.8 Valence-band offset

In the case of an unstrained system the valence-band offset (VBO) has a meaning analogous to that in the conduction band: V here shows the valence-band edge (of both the HH and

LH branches) in any layer. This is precisely how it enters the Hamiltonian. The VBOs at heterointerfaces are generally available from the literature, but there is a considerable amount of scatter in these data. Literature values $V = \Delta V_{\mathrm{VB}}$ for several common unstrained semiconductor systems are given in Appendix A.

The case of strained structures is more complicated, and requires some care in using the available data on VBOs. The potential V, which should be inserted into the Hamiltonian, is *not* any of the band edges defined by equation (14.26), because these expressions already include the effects of strain and cannot be filtered out of the Hamiltonian. Instead, V in the Hamiltonian is the potential before the addition of strain.

In many cases data can be found on the so-called average valence-band energy E_{av}, which is the weighted mean of the three valence bands. In an unstrained material, where HH and LH band edges are degenerate, the weighted mean (in the inverted energy picture!) is $\Delta_{\mathrm{SO}}/3$ above the valence-band edge. Alternatively, the interface of two semiconductors may be characterised by the average VBO, ΔE_{av}, which is the difference of the E_{av}s in the two materials. Since Δ_{SO} is material dependent, the ΔE_{av} is not the same as the VBO, even in an unstrained material. It has been established that ΔE_{av} is roughly constant with strain, and can therefore be used as a single parameter to describe the interface.

If a strained semiconductor layer is grown on an unstrained substrate (made of a different material), and the values of ΔE_{av} and Δ_{SO} in the epitaxial layer material are known, then the difference $\Delta E_{\mathrm{av}} - \Delta_{\mathrm{SO}}/3$ clearly gives the energy of the HH/LH valence-band edge in this material *without* strain, measured from the E_{av} value in the substrate. In the same manner the valence-band edge can be obtained in all layers of a multilayer structure, measured from the same reference point, regardless of whether a particular layer is in direct contact with the substrate or not. This is the potential that should be used in the Hamiltonian. It can be subtracted from any desired reference energy, if it is preferred to have output energies measured from that point. Since the valence-band edge in a layer before strain has no physical significance in the strained system, one reasonable choice for a reference point might be the lowest valence band edge in the quantum well—either HH or LH, whichever came out to be lower (if there are different quantum wells, the deepest one could be chosen). Another choice might be the valence-band edge in the substrate, which is unstrained, and still has its HH and LH band edges degenerate.

In some cases, for a particular material interface and particular strain conditions, data may be quoted such as 'AB on CD has a VBO of ΔE_v (eV)'. This means that strained material AB grown on unstrained substrate CD has such an offset between its valence-band edge (either HH or LH, whichever is the lower in this case) and the valence-band edge in the substrate. To use this information for a calculation within the same material system, but under different strain conditions, it will be necessary to find whether the HH or LH band in the strained AB material makes its valence-band edge (which depends on the sign of the strain), then find P_ϵ and Q_ϵ and then use the appropriate expression from equation (14.26) to get the AB valence-band edge without strain.

14.9 The layer (transfer matrix) method

In order to describe the procedure of finding the quantised state energies and wave functions, the block-diagonal form of the Hamiltonian will be used, i.e. attention will be focused on just

one of the 3×3 blocks. However, the method is straightforwardly applicable to any other size of the Hamiltonian, e.g. the 2×2, or one of the non-block-diagonalised 6×6 or 4×4 forms, or in fact any other form (see [7]).

Consider a structure modulated along one dimension (i.e. a quantum well) with arbitrarily varying material composition and potential (see Figs. 1.9–1.12 for a few examples). For the purpose of finding its bound states, or the tunnelling probability, the heterostructure of interest is subdivided into a number (N_z, number of coordinate points) of thin layers, and within each layer the potential (which includes any self-consistent potential, if such a calculation was performed) is taken to be constant, as are the values of the Luttinger parameters (note that, under these conditions, $S = \Sigma$ and $C = 0$ in equations (14.29) and (14.30)). If the structure is step-graded, comprising some number of layers of different material composition and width (perhaps like the structure in Fig. 3.16), and provided any continuously varying potentials (e.g. self-consistent potential) are absent, the computational layers in this calculation coincide with actual material layers, and need not be very thin.

Similarly to the case of quantised states of electrons, the aim is to describe the wave function of a quantised state in terms of its form within each layer. These are then joined at the interfaces using appropriate boundary conditions. Within such an approach the need is not to generate all the energy eigenstates of the bulk Hamiltonian for a specified wave vector, but rather all the solutions that correspond with a definite energy, i.e. the complex band structure has to be found.

In order to find the complex band structure in the valence band, consider a structure grown in the [001] direction, which it is convention to define as the z-axis. Since the structure composition and the potential are modulated (varied) along z, but are constant along the x- and y-axes, the wave function in the x–y plane must behave like a plane wave, hence k_x and k_y must be real, while k_z is arbitrary. The Hamiltonian is written as (the factor $\hbar^2/2m_0$ is absorbed into the γ parameters to simplify notation):

$$\mathsf{H} = \mathsf{H}_2 k_z^2 + \mathsf{H}_1 k_z + \mathsf{H}_0 \tag{14.31}$$

where H_0, H_1 and H_2 are the 3×3 matrices that are associated with the corresponding powers of the wave vector k_z, i.e.

$$\mathsf{H}_2 = \begin{pmatrix} (\gamma_1 - 2\gamma_2) & 0 & 0 \\ 0 & (\gamma_1 + 2\gamma_2) & -2\sqrt{2}\gamma_2 \\ 0 & -2\sqrt{2}\gamma_2 & \gamma_1 \end{pmatrix}$$

$$\mathsf{H}_1 = \begin{pmatrix} 0 & \mp i2\sqrt{3}\gamma_3 k_{||} & \pm i\sqrt{6}\gamma_3 k_{||} \\ \pm i2\sqrt{3}\gamma_3 k_{||} & 0 & \mp i3\sqrt{2}\gamma_3 k_{||} \\ \mp i\sqrt{6}\gamma_3 k_{||} & \pm i3\sqrt{2}\gamma_3 k_{||} & 0 \end{pmatrix}$$

$$\mathsf{H}_0 = \begin{pmatrix} (\gamma_1 + \gamma_2)k_{||}^2 + P_\epsilon + Q_\epsilon + V & -\sqrt{3}\gamma_\phi k_{||}^2 & -\sqrt{6}\gamma_\phi k_{||}^2 \\ -\sqrt{3}\gamma_\phi k_{||}^2 & (\gamma_1 - \gamma_2)k_{||}^2 + P_\epsilon - Q_\epsilon + V & \sqrt{2}(\gamma_2 k_{||}^2 + Q_\epsilon) \\ -\sqrt{6}\gamma_\phi k_{||}^2 & \sqrt{2}(\gamma_2 k_{||}^2 + Q_\epsilon) & \gamma_1 k_{||}^2 + P_\epsilon + \Delta_{\mathrm{SO}} + V \end{pmatrix} \tag{14.32}$$

At any specified energy E, the values of the complex wave vector k_z may be viewed as eigenvalues of the 3×3 nonlinear eigenvalue problem $[\mathsf{H}(k_z) - E\mathrm{I}]\,\mathbf{F} = 0$, where $\mathbf{F}$ is the eigenfunction vector of length 3. This eigenproblem is nonlinear because eigenvalues (k_z, not E!) appear in powers of both 1 and 2, in contrast to the standard linear eigenproblem where the energy appears only linearly on the diagonal. The nonlinear problem can be solved by a trick that converts it to a double-sized, 6×6 *linear* eigenvalue problem:

$$\begin{pmatrix} \mathsf{0} & \mathrm{I} \\ -\mathsf{H}_2^{-1}(\mathsf{H}_0 - E\mathrm{I}) & -\mathsf{H}_2^{-1}\mathsf{H}_1 \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ k_z\mathbf{u} \end{pmatrix} = k_z \begin{pmatrix} \mathbf{u} \\ k_z\mathbf{u} \end{pmatrix} \tag{14.33}$$

where $\mathsf{0}$ and I are the 3×3 null and identity matrices, respectively. Note that the first row in equation (14.33) is just an identity. The solution of this non-Hermitian but linear matrix eigenproblem, by standard diagonalisation routines [8], delivers six generally complex-valued wave vectors k_z, and the corresponding eigenstates. These states, denoted as $\mathbf{u}$, are expressed in the basis F_1, F_2, F_3 for the upper block. The $\mathbf{u}$-states are thus linear combinations of F_1, F_2, F_3 that, at energy E, behave as plane waves, i.e. $\exp(ik_z z)$, within a layer. Note that the first three components of an eigenvector of equation (14.33) are the basis state amplitudes, and the other three, *when multiplied by* i, will be their derivatives. In further considerations the notation $\mathbf{u}$ will denote the vector of length 6, containing both amplitudes and *derivatives*.

For convenience, these states may then be divided into two groups as follows. A state which has purely real k_z may be classified according to the sign of k_z:[2] if it is positive the state is 'forward', otherwise it is 'backward'. If k_z is complex, including purely imaginary, it is classified as forward if the wave function decays to the right, otherwise it is backward. There are three forward and three backward states in any one layer, and any wave function in the structure may, in this layer, be written as a linear combination of these six states. The vector containing the three coefficients of forward states will be denoted as $\mathbf{a}_j$, and that for backward states as $\mathbf{b}_j$, where the subscript j denotes that these correspond to layer j in the structure. In a vector of length 6, which gives the amplitudes of these states in a wave function at a particular energy, the choice is to write $\mathbf{a}_j$ first and $\mathbf{b}_j$ below it.

Having the wave function at the beginning (left-hand side) of the jth layer, written in the basis of eigen-k_z states u, it may be propagated to the beginning of the next layer, by first multiplying the state vector by the corresponding exponentials, i.e. by left-multiplying the state vector with the diagonal matrix $\mathsf{D}_j = \mathrm{diag}(\ldots, \exp(ik_{z\,j}d_j), \ldots)$, where d_j is the width of the jth layer, and then using the boundary conditions (the interface matching matrix) to propagate the wave function just across the interface. However, the interface matching matrix is written in the $\mathbf{F}$-basis (equation (14.1)), as noted above, and not in the eigen-k_z states basis. It relates the wave function amplitudes and derivatives, written in the $\mathbf{F}$-basis, and is obtained from equation (14.29) in our case (or, if not using the block-diagonal form, the corresponding form would be obtained from equation (14.27)). It requires that:

$$\begin{pmatrix} \mathrm{I} & \mathsf{0} \\ \mathsf{B}_1 & \mathsf{B}_2 \end{pmatrix} \begin{pmatrix} \mathbf{F} \\ \mathbf{F}' \end{pmatrix} = \mathrm{const} \tag{14.34}$$

[2] Strictly speaking, it should be tested for the current it carries, where the current density operator is $(2\mathsf{H}_2 k_z + \mathsf{H}_1)/\hbar$, but for the purpose here the sign test suffices.

across the interface, where:

$$\mathsf{B}_1 = \begin{pmatrix} 0 & \pm 2\sqrt{3}\pi k_{||} & \mp\sqrt{6}\pi k_{||} \\ \mp 2\sqrt{3}(\sigma-\delta)k_{||} & \pm 2(\sigma-\delta-\pi)k_{||} & \pm\sqrt{2}(2\sigma-2\delta+\pi)k_{||} \\ \pm\sqrt{6}(\sigma-\delta)k_{||} & \mp\sqrt{2}(\sigma-\delta+2\pi)k_{||} & \mp 2(\sigma-\delta-\pi)k_{||} \end{pmatrix}$$

$$\mathsf{B}_2 = \begin{pmatrix} \gamma_1 - 2\gamma_2 & 0 & 0 \\ 0 & \gamma_1 + 2\gamma_2 & -2\sqrt{2}\gamma_2 \\ 0 & -2\sqrt{2}\gamma_2 & \gamma_1 \end{pmatrix} \tag{14.35}$$

and:

$$\begin{pmatrix} \mathbf{F} \\ \mathbf{F}' \end{pmatrix} = \begin{pmatrix} F_1 \\ F_2 \\ F_3 \\ F_1' \\ F_2' \\ F_3' \end{pmatrix} \tag{14.36}$$

where the '$+$' sign applies to the upper block, and the '$-$' would apply to the lower block (with F_4, F_5, F_6 functions) when required. The matrix in equation (14.34) will be denoted as M_j (the interface matrix, corresponding to the layer j). From $\mathsf{M}_{j+1}(\mathbf{F}\ \ \mathbf{F}')_{j+1}^T = \mathsf{M}_j(\mathbf{F}\ \ \mathbf{F}')_j^T$ it follows that the matrix $\mathsf{M}_{j+1}^{-1}\mathsf{M}_j$ transfers a state written in the $\mathbf{F}$-basis from the left-hand side into the right-hand side of the interface of layers j and $j+1$.

A wave function written in the $\mathbf{u}$-basis, can be re-expressed in the $\mathbf{F}$-basis (amplitudes and derivatives) by multiplying it with the matrix U, made by stacking side-by-side all the eigen-k_z states $u_1, \ldots, u_6$ (i.e. eigenvectors of equation (14.33) in which the lower components have been turned into derivatives, as noted above):

$$\mathsf{U} = (u_1|u_2|\ldots|u_6)\,, \quad \mathbf{F} = \mathsf{U}\mathbf{u}, \quad \mathbf{u} = \mathsf{U}^{-1}\mathbf{F} \tag{14.37}$$

Therefore, the transfer matrix $\mathsf{T}^{j,j+1}$, relating the wave functions at the beginning of layer j and the beginning of layer $j+1$ (with wave functions written in the $\mathbf{u}$-basis) is given by:

$$\mathsf{T}^{j,j+1} = \mathsf{U}_{j+1}^{-1}\mathsf{M}_{j+1}^{-1}\mathsf{M}_j\mathsf{U}_j\mathsf{D}_j, \quad \begin{pmatrix} \mathbf{a}_{j+1} \\ \mathbf{b}_{j+1} \end{pmatrix} = \mathsf{T}^{j,j+1}\begin{pmatrix} \mathbf{a}_j \\ \mathbf{b}_j \end{pmatrix} \tag{14.38}$$

To give a more verbose description of equation (14.38): taking a state at the beginning of layer j written in the $\mathbf{u}$-basis, first propagate it to the end of that layer, but still inside it (D_j), then express it in the $\mathbf{F}$-basis (U_j), then transfer it across the interface into the beginning of layer $j+1$ ($\mathsf{M}_{j+1}^{-1}\mathsf{M}_j$), and finally express it back in the $\mathbf{u}$-basis of layer $j+1$ (U_{j+1}^{-1}). The total transfer matrix of a structure consisting of N layers is then:

$$\mathsf{T}^{0,N} = \mathsf{T}^{N-1,N}\ldots\mathsf{T}^{1,2}\mathsf{T}^{0,1} \tag{14.39}$$

How can the bound states of a system be found if the transfer matrix is known? First, note that it can be written in terms of its four 3×3 blocks as:

$$\mathsf{T} = \begin{pmatrix} \mathsf{T}_{11} & \mathsf{T}_{12} \\ \mathsf{T}_{21} & \mathsf{T}_{22} \end{pmatrix} \tag{14.40}$$

so that:

$$\begin{pmatrix} \mathbf{a}_N \\ \mathbf{b}_N \end{pmatrix} = \begin{pmatrix} \mathsf{T}_{11} & \mathsf{T}_{12} \\ \mathsf{T}_{21} & \mathsf{T}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{a}_0 \\ \mathbf{b}_0 \end{pmatrix} \tag{14.41}$$

This is similar to the familiar case of electrons in the conduction band, except that the transfer matrix there is of size 2×2 and its elements are numbers, while here these numbers are replaced by matrix blocks.

Now assume that, for a particular in-plane wave vector and energy, the complex bandstructure calculation for the outermost two layers of a quantum well delivers a pair of purely real-valued k_z. In this case (which is analogous to the case of the energy being above the barrier for electrons) there are no bound states. However, if all the k_zs in some energy range turn out to be complex-valued or simply imaginary, there may be bound states because the wave function composed of such states can be made to decay away from the quantum well region. To find these energies and wave functions, it must be noted that, on the left-hand side, a bound state (whose wave function can be normalised to unity) must have $\mathbf{a}_0 = \mathbf{0}$, while $\mathbf{b}_0$ is non-zero. Therefore, from equation (14.41), on the right-hand side, the wave function has the component $\mathbf{a}_N = \mathsf{T}_{12}\mathbf{b}_0$, which decays further to the right, and the component $\mathbf{b}_N = \mathsf{T}_{22}\mathbf{b}_0$ which grows to the right. However, this latter term cannot exist in a bound-state wave function, hence it must be the case that:

$$\mathsf{T}_{22}\mathbf{b}_0 = \mathbf{0} \tag{14.42}$$

For electronic bound states, where T_{22} and $\mathbf{b}_0$ are both scalars, the method would be to search for energies that make $T_{22} = 0$, in which case b_0 is arbitrary, and eventually determined from the normalisation of the wave function. For holes, equation (14.42) is, in fact, a homogeneous system of linear equations in the components of $\mathbf{b}_0$. To have a non-trivial solution, the determinant of the matrix T_{22} must be equal to zero, and this will happen only for some particular values of the energy. Therefore, when searching for bound states, the energy is varied and the value of $\det|\mathsf{T}_{22}(E)| = 0$ is monitored (note that this determinant is generally a complex number). Upon finding such value(s) of the energy E, the system of equations for $\mathbf{b}_0$ is solved, which can be done to within an arbitrary multiplier (its value is eventually determined from normalisation). Finally, starting with this $\mathbf{b}_0$, and $\mathbf{a}_0 = \mathbf{0}$ and using the transfer matrix once again, the wave function for each particular bound state can be generated.

In the energy range where at least one of the wave vectors k_z is real, there are no bound states, but the transmission or reflection coefficient of the structure can be calculated. In this case it would be necessary to specify the incident state (all its components, not just one amplitude) for which the transmission is sought.

Despite its more complicated procedure, as described above, the layer method is usually much faster than any direct numerical procedure for finding the bound states. This is because it handles only small matrices, while direct diagonalisation uses very large matrices which, in the case of holes, are not tridiagonal. Furthermore, this was a nice illustration of the application of the transfer matrix method to more complex systems, which is why it was considered in detail.

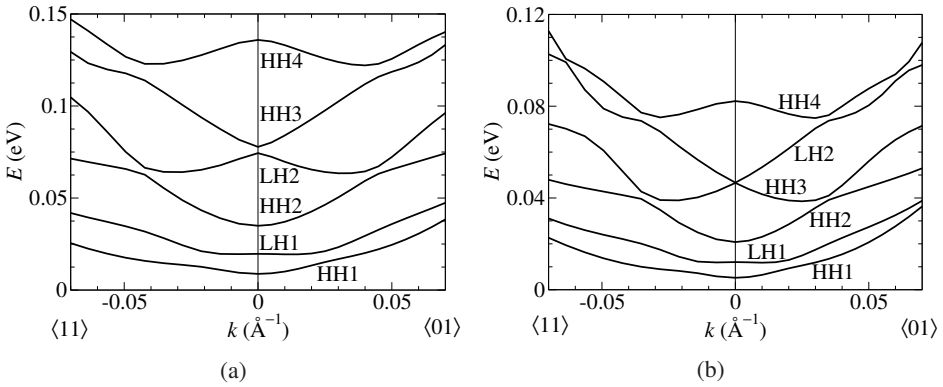


Figure 14.2: The dispersion of the first six hole subbands in (a) an 85 Å wide and (b) a 113 Å wide GaAs quantum well surrounded by $Ga_{0.5}Al_{0.5}As$ barriers. The energy is measured from the valence-band edge in the well

14.10 Quantum well subbands

In this section examples of hole subbands will be given, calculated for GaAs quantum wells embedded in the $Ga_{0.5}Al_{0.5}As$ bulk alloy. The valence-band edge of GaAs is lower than that of the alloy (in the inverted energy picture), so the GaAs layer forms the quantum well and $Ga_{0.5}Al_{0.5}As$ the barrier. GaAs and AlAs have such similar values of the lattice constant that the structure is taken to be lattice-matched, i.e. unstrained. The material parameters were taken from Table 14.1.

Subband dispersion curves (E versus k dependence) in two structures with different quantum well widths are given in Fig. 14.2. The wave vector in both examples was varied along two different directions in the k_x–k_y plane: the $\langle 01 \rangle$ direction ($k_x \neq 0$ and $k_y = 0$, or vice versa), and the $\langle 11 \rangle$ direction ($k_x = k_y$). It is immediately apparent that the dispersion is quite anisotropic.

The subbands are denoted as HH or LH according to their composition at $k_x = k_y = 0$. For finite in-plane wave vectors, particularly in the outer parts of Fig. 14.2, the mixing of $|3/2, \pm 3/2\rangle$ and $|3/2, \pm 1/2\rangle$ states becomes so strong that it would be difficult to tell, by looking at the wave function composition alone, whether a subband is of HH or LH type. The label HH, for instance, then just means that this state originated as HH at the Γ-point.

The wave-vector-dependent mixing of bulk bands is the reason for highly non-parabolic subband dispersion along any particular direction, even with occasional reversal ($\partial E/\partial k$ becoming negative, which can be interpreted as a negative mass). The ordering of the first few subbands is the same for both well widths: HH1 is the lowest, then LH1, then HH2 (this remains so for any well width), but differences appear for higher subbands. It is interesting to point out that, in the (relatively low) energy range shown in Fig. 14.2, the contribution of the bulk SO band (i.e. the $|1/2, \pm 1/2\rangle$ basis states) to the subband wave functions is very small. From less than 1% at $k_x = k_y = 0$ in the LH subband, it increases but stays typically between less than 10% and 15% in either the HH or the LH subbands almost throughout the

range shown in Fig. 14.2.

14.11 The influence of strain

Calculation of the hole subband structure in strained systems proceeds by the same method as in the unstrained case, the only difference being the presence of strain terms in the Hamiltonian. In this section the effects of strain on states in quantum wells will be discussed.

Consider a compressively strained well between two tensile strained barriers. An example would be a $Si_{1-x}Ge_x$ layer embedded in Si, grown on a substrate with a composition between the two (this is a common configuration in order to achieve strain balance). The valence-band edge in the SiGe alloy is lower than in Si (in the inverted energy picture), so the former is the well and the latter the barrier. The larger lattice constant of Ge and, hence, the compressive strain in the $Si_{1-x}Ge_x$ layer imply that the HH band edge will decrease from the unstrained value, while the LH band edge would increase (note again the use of the inverted energy picture). The opposite applies in the tensile strained Si barriers. Therefore, the quantum well becomes deeper for heavy holes and shallower for light holes. This is a direct consequence of the uniaxial strain component, which splits the HH and LH band edges apart.

The hydrostatic component also influences the barrier heights, because the hydrostatic deformation potential has different values in the two materials, but this does not affect the bound state energies very much. If the barriers are relatively high, the bound state energies *relative to the band edge* mostly depend on well width, and the strain-induced shift of the band edge (HH or LH) will essentially 'drag' with it all of 'its own' bound states (however, due to mixing of the bulk states, the meaning of this is somewhat vague for non-zero values of $k_{||}$). In the unstrained case the lowest HH state is below the lowest LH state, therefore compressive strain will displace the two sets of states, leaving HH as the lowest.

In the opposite case of a well under tensile strain, the valence-band edges behave in exactly the opposite manner, and this can lead to the unusual situation of the lowest LH state sinking below the lowest HH state. In the SiGe system it is impossible for the LH subband to be the lowest as this would require a tensilely strained well, and in this case the well (the layer with the larger Ge content) is always compressively strained. However, there are other materials (based on III–V alloys) in which the well can be put under tensile strain through an appropriate choice of parameters, and the LH1 subband can become the lowest in the system.

14.12 Strained quantum well subbands

As an example of a strained quantum well, take a $Si_{0.6}Ge_{0.4}$ layer embedded between Si layers, grown on a $Si_{0.7}Ge_{0.3}$ substrate. The calculated subband dispersion in two structures with different well widths, using the material parameters from Table 14.1 and the value $\Delta E_{\text{av}} = 0.56\,\text{eV}$, is shown in Fig. 14.3. Along with the anisotropic dispersion, just as in the case of the unstrained structure in Fig. 14.2, notice that the ordering of subbands now depends on the well width: it is possible to get two (or even more) HH subbands below the lowest LH subband. This occurs because of the strain-induced displacement of the HH and LH band edges. Another interesting feature in Fig. 14.3(b) is the dispersion of the HH2 subband: starting from the zone centre, its energy first decreases before acquiring the expected increase

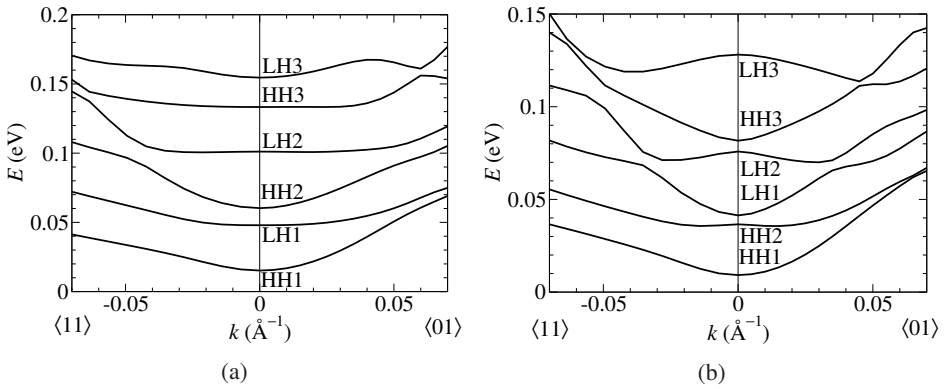


Figure 14.3: The dispersion of the first six hole subbands in (a) an 83 Å wide and (b) a 111 Å wide $Si_{0.6}Ge_{0.4}$ quantum well embedded in Si barriers, grown on a $Si_{0.7}Ge_{0.3}$ substrate. The energy is measured from the HH band edge in the strained well material

with the wave vector, i.e. it shows an inverted mass feature, which is brought about by mixing with the LH1 subband lying just above it.

In contrast to the GaAs/AlGaAs system, in this example the SO band contributes significantly to the composition of all subbands. At the Γ-point its contribution to the LH subbands is of the order of $\sim 10\%$ (this is a consequence of strain alone), and then increases significantly for larger in-plane wave vectors for all the subbands.

14.13 Direct numerical methods

An alternative method of finding the quantised states is to solve equation (14.27) or (14.29) when written for the whole structure. One possibility is to use the finite-difference approximation, which was introduced for approximating derivatives in the effective mass Schrödinger equation of Chapter 3. In the valence band the wave function has more than a single component, hence finite differences never lead to a tridiagonal matrix, even for one-dimensional problems (quantum wells), but rather to a band matrix having non-zero elements along the main and some number of remote diagonals. In the usual implementation the finite-difference method implicitly assumes hard wall (box) boundary conditions, as in Fig. 3.6.

Another possibility is to use a plane wave method and expand the wave function components in a Fourier series, hence implicitly assuming the structure to be a superlattice, even though it may be non-periodic—for example, a single quantum well embedded in thick barriers. To prevent interaction between the adjacent wells in implicitly existing neighbouring periods, care has to be taken to ensure that the barriers are thick enough. The plane wave method is explained in the chapter on empirical pseudo-potentials (Chapter 15), and will not be elaborated on here in any more detail.

The advantage of these methods is that they may be used equally well for structures of higher dimensionality (i.e. quantum wires and quantum dots) with no essential complications

in the formulation. However, these methods require considerable amounts of computer memory in order to store the necessary matrices, especially in multidimensional systems, and such matrices also require long computation time for diagonalisation. An alternative is to employ a suitable set of basis functions that are adapted to the problem at hand, in the sense that using a relatively small number of these functions suffices for good convergence.

Exercises

(1). Using equation (14.9), find expressions for the effective masses of the heavy-hole (HH) and light-hole (LH) bands if the direction of the k-vector is [001], [110] and [111]. Calculate these masses for GaAs.

(2). Starting with the upper 2×2 block of the Hamiltonian (equation (14.19)), find the quantised state energies for heavy and light holes in an infinitely deep quantum well in the case where the in-plane wave vector $k_{\|} = 0$. Then, using perturbation theory up to second order, and considering all terms with $k_{\|}$ in the Hamiltonian as a perturbation, derive an expression for the in-plane effective mass of the heavy- (HH) and light-hole (LH) states in such a quantum well.

For a GaAs well of width $L = 10\,\mathrm{nm}$, find the energies of the lowest four HH and lowest four LH states. By keeping only the most dominant term in the perturbation expressions, calculate the in-plane effective masses of the two lowest HH and LH states.

References

[1] J. M. Luttinger and W. Kohn, 'Motion of electrons and holes in perturbed periodic fields', *Phys. Rev.*, **97**:869, 1955.

[2] C. R. Pidgeon and R. N. Brown, 'Interband magneto-absorption and Faraday rotation in InSb', *Phys. Rev.*, **146**:575, 1966.

[3] B. A. Foreman, 'Effective-mass Hamiltonian and boundary conditions for the valence bands of semiconductor microstructures', *Phys. Rev. B*, **48**:4964, 1993.

[4] I. Vurgaftman, J. R. Meyer, and L. R. Ram-Mohan, 'Band parameters for III–V compound semiconductors and their alloys', *J. Appl. Phys.*, **89**(11):5815–5875, 2001.

[5] A. Kahan, M. Chi, and L. Friedman, 'Infrared transitions in strained-layer Ge_xSi_{1-x}/Si', *J. Appl. Phys.*, **75**(12):8012–8021, 1994.

[6] C. Y. P. Chao and S. L. Chuang, 'Spin-orbit-coupling effects on the valence-band structure of strained semiconductor quantum wells', *Phys. Rev. B*, **46**:4110, 1992.

[7] B. Chen, M. Lazzouni, and L. R. Ram-Mohan, 'Diagonal representation for the transfer matrix method for obtaining electronic energy levels in layered semiconductor heterostructures', *Phys. Rev. B*, **45**:1204, 1992.

[8] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, Third edition, 1999.

15

Empirical pseudo-potential band structure

15.1 Principles and approximations

The envelope function and effective mass approximations, which, as stated earlier, can be thought of as an approximation to the band structure of a crystal, rather than to the quantum mechanics, are very successful theories, which have allowed many of the most fundamental properties of semiconductors and their heterostructures to be explained. However, it is clear that it is at least worthwhile considering more complex, perhaps more fundamental, models to see if they can offer more insight. In particular, given the approximations that the 'single band' effective mass and envelope function models made to the crystal potential, an obvious improvement would be to consider the potential microscopically, i.e. including the potentials of the atoms individually.

The complexity of such a procedure—solving an exact solid—is illustrated by the form of the complete Hamiltonian:

$$\mathcal{H} = \mathcal{H}_{\text{electrons}} + \mathcal{H}_{\text{nuclei}} + \mathcal{H}_{\text{electrons–nuclei}} \tag{15.1}$$

Here

$$\mathcal{H}_{\text{electrons}} = \sum_{\mu} \left(-\frac{\hbar^2}{2m_0}\nabla_{\mu}^2 + \sum_{\lambda<\mu} \frac{1}{4\pi\epsilon_0} \frac{e^2}{|\mathbf{r}_{\lambda} - \mathbf{r}_{\mu}|} \right) \tag{15.2}$$

in which $\mathbf{r}_{\mu}$ are the positions of the electrons and m_0 is the mass. In addition:

$$\mathcal{H}_{\text{nuclei}} = \sum_{\nu} \left(-\frac{\hbar^2}{2M_{\nu}}\nabla_{\nu}^2 + \sum_{\lambda<\nu} \frac{1}{4\pi\epsilon_0} \frac{Z_{\lambda}Z_{\nu}e^2}{|\mathbf{R}_{\lambda} - \mathbf{R}_{\nu}|} \right) \tag{15.3}$$

where $\mathbf{R}_{\nu}$ are the positions, Z_{ν} the atomic numbers and M_{ν} the masses of the nuclei. Finally:

$$\mathcal{H}_{\text{electrons–nuclei}} = -\sum_{\mu,\nu} \frac{1}{4\pi\epsilon_0} \frac{Z_{\nu}e^2}{|\mathbf{R}_{\nu} - \mathbf{r}_{\mu}|} \tag{15.4}$$

In a typical macroscopic sample of semiconducting crystal, there are a lot of nuclei, and correspondingly a lot of electrons! In fact, for Si, there are about 5×10^{22} atoms per cubic

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

centimetre. The problem, as stated above, is therefore insolvable. There are, however, some approximations [1, 2] that can be made, which have proved to be acceptably accurate and thus lead to a more manageable problem.

(1). Assuming that the electrons below the outer shell are tightly bound to the nucleus simplifies the system slightly—the atoms can then be treated as single entities and hence the number of particles is reduced. In addition, these inner-shell electrons screen the outer valence electrons from the central nuclear charge, which has the effect of smoothing out the potential of what would otherwise be a rapidly varying term. Thinking ahead slightly, the smoother the potential, the smaller the number of terms that may be required if the electron wave functions were to be constructed from some Fourier series.

(2). The *adiabatic approximation* assumes that a change in the coordinates of a nucleus passes no energy to the electrons, i.e. the electrons respond adiabatically, which then allows the decoupling of the motions of the nuclei and the electrons.

(3). The *independent electron approximation* removes the complications of the electron–electron interactions and replaces them with a time-averaged potential.

These approximations allow the system, i.e. the electron wave functions that represent the complete self-interacting electron cloud within the 'plasma' of the crystal, to be collapsed down to a one-electron problem with the following Hamiltonian:

$$\mathcal{H} = -\frac{\hbar^2}{2m_0}\nabla^2 + V_{\mathrm{c}} \tag{15.5}$$

where V_{c} is some, as yet undetermined, crystal potential that represents not only the interaction between the electrons and the nuclei of the atoms that constitute the lattice, but also the interaction between the electrons themselves.

15.2 Elemental band structure calculation

It is expected *a priori* that the solutions obtained from a model that will rely on the periodicity of the potential will be dependent on the electron momentum **k**, since it is known that electric current flows in semiconductors and from experiences with the Kronig–Penney model. In addition, as more than one solution may be obtained, then the electron wave function may be most generally labelled as $\psi_{n,\mathbf{k}}$. Given these features, using the time-independent (as the potential is static) Schrödinger equation gives:

$$\mathcal{H}\psi_{n,\mathbf{k}} = E_{n,\mathbf{k}}\psi_{n,\mathbf{k}} \tag{15.6}$$

Being influenced by all the previous success in expanding periodic functions in terms of linear combination of a set of basis functions, where each has the periodicity of the system, and by utilising the idea of Bloch ([3], p. 133), for example, it is tempting to expand the wave function $\psi_{n,\mathbf{k}}$ in terms of the complete orthonormal set of plane waves:

$$u_{\mathbf{G},\mathbf{k}} = \frac{1}{\sqrt{\Omega}}\mathrm{e}^{i(\mathbf{G}+\mathbf{k})\bullet\mathbf{r}} \tag{15.7}$$

and then:

$$\psi_{n,\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} a_{n,\mathbf{k}}(\mathbf{G}) u_{\mathbf{G},\mathbf{k}} \tag{15.8}$$

$$\therefore \psi_{n,\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{\Omega}} \sum_{\mathbf{G}} a_{n,\mathbf{k}}(\mathbf{G}) \mathrm{e}^{i(\mathbf{G}+\mathbf{k})\bullet\mathbf{r}} \tag{15.9}$$

If **R** is a Bravais lattice vector, then:

$$\psi_{n,\mathbf{k}}(\mathbf{r}+\mathbf{R}) = \frac{1}{\sqrt{\Omega}} \sum_{\mathbf{G}} a_{n,\mathbf{k}}(\mathbf{G}) \mathrm{e}^{i(\mathbf{G}+\mathbf{k})\bullet(\mathbf{r}+\mathbf{R})} \tag{15.10}$$

which gives:

$$\psi_{n,\mathbf{k}}(\mathbf{r}+\mathbf{R}) = \frac{1}{\sqrt{\Omega}} \sum_{\mathbf{G}} a_{n,\mathbf{k}}(\mathbf{G}) \mathrm{e}^{i(\mathbf{G}+\mathbf{k})\bullet\mathbf{R}} \mathrm{e}^{i(\mathbf{G}+\mathbf{k})\bullet\mathbf{r}} \tag{15.11}$$

$$\therefore \psi_{n,\mathbf{k}}(\mathbf{r}+\mathbf{R}) = \mathrm{e}^{i\mathbf{k}\bullet\mathbf{R}} \frac{1}{\sqrt{\Omega}} \sum_{\mathbf{G}} a_{n,\mathbf{k}}(\mathbf{G}) \mathrm{e}^{i\mathbf{G}\bullet\mathbf{R}} \mathrm{e}^{i(\mathbf{G}+\mathbf{k})\bullet\mathbf{r}} \tag{15.12}$$

which yields:

$$\psi_{n,\mathbf{k}}(\mathbf{r}+\mathbf{R}) = \mathrm{e}^{i\mathbf{k}\bullet\mathbf{R}} \psi_{n,\mathbf{k}}(\mathbf{r}) \tag{15.13}$$

i.e. Bloch's theorem, provided that a set of wave vectors **G** can be found for the original expansion, which satisfy:

$$\mathbf{G}\bullet\mathbf{R} = 2\pi n, \qquad \text{where } n \in \mathcal{Z} \tag{15.14}$$

Of course, such a set of vectors does exist—they are merely the reciprocal lattice vectors describing the periodicity of the lattice (see Chapter 1).

Therefore, the proposed expansion of the wave function in terms of a linear combination of plane waves satisfies Bloch's theorem and thus it is worthwhile proceeding further in order to deduce the consequences. Hence, substituting the expansion for $\psi_{n,\mathbf{k}}$ given in equation (15.9) into the Schrödinger equation (equation (15.6)):

$$\mathcal{H} \sum_{\mathbf{G}} a_{n,\mathbf{k}}(\mathbf{G}) u_{\mathbf{G},\mathbf{k}} = E_{n,\mathbf{k}} \sum_{\mathbf{G}} a_{n,\mathbf{k}}(\mathbf{G}) u_{\mathbf{G},\mathbf{k}} \tag{15.15}$$

Utilising the *linearity* of the Hamiltonian, i.e. it acts on each $u_{\mathbf{G},\mathbf{k}}$ in turn, equation (15.15) can be written:

$$\sum_{\mathbf{G}} a_{n,\mathbf{k}}(\mathbf{G}) \mathcal{H} u_{\mathbf{G},\mathbf{k}} = E_{n,\mathbf{k}} \sum_{\mathbf{G}} a_{n,\mathbf{k}}(\mathbf{G}) u_{\mathbf{G},\mathbf{k}} \tag{15.16}$$

Multiplying through by a particular plane wave $u^*_{\mathbf{G}',\mathbf{k}}$ and integrating over all space:

$$\sum_{\mathbf{G}} a_{n,\mathbf{k}}(\mathbf{G}) \int u^*_{\mathbf{G}',\mathbf{k}} \mathcal{H} u_{\mathbf{G},\mathbf{k}} \, \mathrm{d}\tau = \sum_{\mathbf{G}} a_{n,\mathbf{k}}(\mathbf{G}) E_{n,\mathbf{k}} \int u^*_{\mathbf{G}',\mathbf{k}} u_{\mathbf{G},\mathbf{k}} \, \mathrm{d}\tau \tag{15.17}$$

Writing:

$$H_{\mathbf{G}',\mathbf{G}} = \int u^*_{\mathbf{G}',\mathbf{k}} \mathcal{H} u_{\mathbf{G},\mathbf{k}} \, \mathrm{d}\tau \tag{15.18}$$

and using the orthonormality property of the basis set, i.e.

$$\int u^*_{\mathbf{G}',\mathbf{k}} u_{\mathbf{G},\mathbf{k}} \, \mathrm{d}\tau = \delta_{\mathbf{G}',\mathbf{G}} \tag{15.19}$$

equation (15.17) then becomes:

$$\sum_{\mathbf{G}} a_{n,\mathbf{k}}(\mathbf{G}) H_{\mathbf{G}',\mathbf{G}} = \sum_{\mathbf{G}} a_{n,\mathbf{k}}(\mathbf{G}) E_{n,\mathbf{k}} \delta_{\mathbf{G}',\mathbf{G}} \tag{15.20}$$

This expression can be written more succinctly as a matrix equation, where each plane wave, $\mathbf{G}'$ corresponds to a different row of the matrix. The left-hand side of equation (15.20) corresponds to a multiplication between a row of matrix H and a column vector $\mathbf{a}_{n,\mathbf{k}}$. The Kronecker delta function on the right-hand side reduces it to a multiplication between a scalar value $E_{n,\mathbf{k}}$ and an element of the column vector $\mathbf{a}_{n,\mathbf{k}}$. Therefore, the complete matrix form of equation (15.20) is:

$$\mathsf{H}\mathbf{a}_{n,\mathbf{k}} = E_{n,\mathbf{k}} \mathbf{a}_{n,\mathbf{k}} \tag{15.21}$$

An equation such as this exists for each value of the electron wave vector $\mathbf{k}$. The solution is therefore reduced to finding the eigenvalues and eigenvectors of the square matrix H (whose matrix elements are denoted $H_{\mathbf{G}',\mathbf{G}}$). In principle, the matrix is of infinite order, but in practice the expansion is limited to a finite set of N plane waves $\mathbf{G}$, and hence the problem can be solved by direct diagonalisation using one of the many computer libraries that are available. This yields N eigenvalues $E_{n,\mathbf{k}}$ (which represent the energies of the nth state at wave vector $\mathbf{k}$) and the corresponding eigenvectors $\mathbf{a}_{n,\mathbf{k}}$, whose elements give the weighting of each plane wave basis state in the expanded wave function of the state.

It remains now to construct the individual matrix elements $H_{\mathbf{G}',\mathbf{G}}$, given by equation (15.18). The Hamiltonian for an electron of rest mass m_0, moving in the crystal potential V_c, is given by:

$$\mathcal{H} = -\frac{\hbar^2}{2m_0}\nabla^2 + V_\mathrm{c} \tag{15.22}$$

Using the form of the plane waves given in equation (15.7), then:

$$\begin{aligned} H_{\mathbf{G}',\mathbf{G}} = \frac{1}{\Omega} \int \mathrm{e}^{-i(\mathbf{G}'+\mathbf{k})\bullet\mathbf{r}} \left(-\frac{\hbar^2}{2m_0}\nabla^2\right) \mathrm{e}^{i(\mathbf{G}+\mathbf{k})\bullet\mathbf{r}} \, \mathrm{d}\tau \\ + \frac{1}{\Omega} \int \mathrm{e}^{-i(\mathbf{G}'+\mathbf{k})\bullet\mathbf{r}} \, V_\mathrm{c} \, \mathrm{e}^{i(\mathbf{G}+\mathbf{k})\bullet\mathbf{r}} \, \mathrm{d}\tau \end{aligned} \tag{15.23}$$

Therefore:

$$\begin{aligned} H_{\mathbf{G}',\mathbf{G}} = \frac{\hbar^2}{2m_0\Omega} \int |\mathbf{G}+\mathbf{k}|^2 \mathbf{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{r}} \, \mathrm{d}\tau \\ + \frac{1}{\Omega} \int \mathrm{e}^{-i(\mathbf{G}'+\mathbf{k})\bullet\mathbf{r}} \, V_\mathrm{c} \, \mathrm{e}^{i(\mathbf{G}+\mathbf{k})\bullet\mathbf{r}} \, \mathrm{d}\tau \end{aligned} \tag{15.24}$$

Now the integral over all space in the first term only has a value if $\mathbf{G} = \mathbf{G}'$, in which case it is equal to the normalisation volume Ω introduced earlier. This can be summarised by using a Kronecker delta, i.e.

$$H_{\mathbf{G}',\mathbf{G}} = \frac{\hbar^2}{2m_0}|\mathbf{G}+\mathbf{k}|^2 \delta_{\mathbf{G},\mathbf{G}'} + V \tag{15.25}$$

where V is the 'potential' given by:

$$V = \frac{1}{\Omega}\int \mathrm{e}^{-i(\mathbf{G}'+\mathbf{k})\bullet\mathbf{r}}\, V_{\mathrm{c}}\, \mathrm{e}^{i(\mathbf{G}+\mathbf{k})\bullet\mathbf{r}}\, \mathrm{d}\tau \tag{15.26}$$

Now the aim of this exercise is to derive a *microscopic* or *atomistic* model of the crystal, so it is clear that the potential V should be written as a sum of some other, as yet undefined, potential that is situated at every atom site $\mathbf{r}_a$, i.e.

$$V_{\mathrm{c}} = \sum_{\mathbf{r}_a} V_a(\mathbf{r}-\mathbf{r}_a) \tag{15.27}$$

Using this form for V_{c} in equation (15.26) gives:

$$V = \frac{1}{\Omega}\int \mathrm{e}^{-i(\mathbf{G}'+\mathbf{k})\bullet\mathbf{r}} \sum_{\mathbf{r}_a} V_a(\mathbf{r}-\mathbf{r}_a)\mathrm{e}^{i(\mathbf{G}+\mathbf{k})\bullet\mathbf{r}}\, \mathrm{d}\tau \tag{15.28}$$

$$\therefore V = \frac{1}{\Omega}\sum_{\mathbf{r}_a}\int V_a(\mathbf{r}-\mathbf{r}_a)\mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{r}}\, \mathrm{d}\tau \tag{15.29}$$

As the origin for the summation over the atoms is undefined, it is possible, and indeed mathematically convenient, to perform the transformation $\mathbf{r} \to \mathbf{r}+\mathbf{r}_a$, then obtain:

$$V = \frac{1}{\Omega}\sum_{\mathbf{r}_a}\int V_a(\mathbf{r})\mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet(\mathbf{r}+\mathbf{r}_a)}\, \mathrm{d}\tau \tag{15.30}$$

$$\therefore V = \frac{1}{\Omega}\sum_{\mathbf{r}_a}\mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{r}_a}\int V_a(\mathbf{r})\mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{r}}\, \mathrm{d}\tau \tag{15.31}$$

The term $\sum_{\mathbf{r}_a}\mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{r}_a}$ is known as the geometrical structure factor S (see, for example, [3], p. 104). Utilising the property discussed earlier that the atomic sites within the crystal are constructed from the Bravais lattice plus the basis, the sum over the atom sites can be replaced with two separate summations, i.e. one over the Bravais lattice points and one over the points in each basis:

$$S = \sum_{\mathbf{R}}\sum_{\mathbf{t}}\mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet(\mathbf{R}+\mathbf{t})} \tag{15.32}$$

where $\mathbf{R}$ is a Bravais lattice vector and $\mathbf{t}$ is a basis vector. Hence:

$$S = \sum_{\mathbf{R}}\mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{R}}\sum_{\mathbf{t}}\mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{t}} \tag{15.33}$$

Now the difference between two reciprocal lattice vectors, i.e. $\mathbf{G}-\mathbf{G}'$, must be another reciprocal lattice vector, $\mathbf{G}''$. In addition, the scalar product of a Bravais lattice vector with a reciprocal lattice vector is just an integral number of 2π, i.e.

$$\mathbf{R}\bullet\mathbf{G}'' = 2\pi n, \qquad \text{where } n \in \mathcal{Z} \tag{15.34}$$

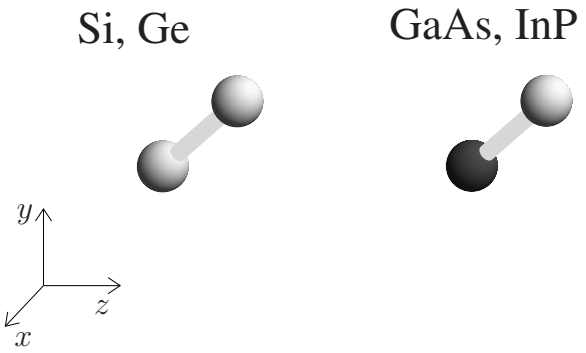


Figure 15.1: The two-atom basis of the diamond structure (left), e.g. Si, Ge; and the zinc blende structure (right), e.g. GaAs, InP

Therefore, $\mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{R}} = 1$, and hence for N Bravais lattice points in the crystal:

$$S = N \sum_{\mathbf{t}} \mathbf{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{t}} \tag{15.35}$$

As discussed earlier in Chapter 1, *most* of the important semiconductors, both elemental and compound, have a face-centred cubic Bravais lattice with a two-atom basis, as shown in Fig. 15.1. The atomic positions within the basis are described by the *basis vector*, $\mathbf{t} = \pm\mathbf{T}$ (say), where:

$$\mathbf{T} = \frac{A_0}{8}(\hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}}) \tag{15.36}$$

Hence, the summation for the structure factor in equation (15.35) contains just two terms, thus giving:

$$S = N\left[\mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{T}} + \mathrm{e}^{-i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{T}}\right] \tag{15.37}$$

and therefore:

$$S = N\left[\cos(\mathbf{G}-\mathbf{G}')\bullet\mathbf{T} + i\sin(\mathbf{G}-\mathbf{G}')\bullet\mathbf{T} + \cos(\mathbf{G}-\mathbf{G}')\bullet\mathbf{T} - i\sin(\mathbf{G}-\mathbf{G}')\bullet\mathbf{T}\right]$$

$$\therefore S = 2N\cos(\mathbf{G}-\mathbf{G}')\bullet\mathbf{T} \tag{15.38}$$

By using this form for the structure factor S, the potential term V, as presented above in equation (15.31), then becomes:

$$V = \frac{2N}{\Omega}\cos(\mathbf{G}-\mathbf{G}')\bullet\mathbf{T}\int V_a(\mathbf{r})\mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{r}}\,\mathrm{d}\tau \tag{15.39}$$

Now the total volume, Ω, divided by the number of Bravais lattice points in the crystal N, is equal to the volume occupied by a single Bravais lattice point, Ω_c (say). Note that, for the

face-centred cubic crystal, the primitive cube of side A_0 (the lattice constant) contains four Bravais lattice points (i.e. eight atoms), and hence:

$$\Omega_{\mathrm{c}} = \frac{A_0^3}{4} \tag{15.40}$$

Furthermore, relabelling $\mathbf{G}' - \mathbf{G}$ as the vector $\mathbf{q}$:

$$V = \frac{2}{\Omega_{\mathrm{c}}} \cos \mathbf{q}_{\bullet} \mathbf{T} \int V_a(\mathbf{r}) \mathrm{e}^{-i\mathbf{q}_{\bullet}\mathbf{r}} \, \mathrm{d}\tau \tag{15.41}$$

The factor:

$$\frac{1}{\Omega_{\mathrm{c}}} \int V_a(\mathbf{r}) \mathrm{e}^{-i\mathbf{q}_{\bullet}\mathbf{r}} \, \mathrm{d}\tau = V_{\mathrm{f}}(q) \tag{15.42}$$

is the Fourier transform of the, as yet undefined, atomic potential V_a, and can be labelled conveniently as $V_{\mathrm{f}}(q)$, where $q = |\mathbf{q}|$. This term is known as the *pseudo-potential form factor* [4].

Table 15.1 The first 65 reciprocal lattice vectors of a face-centred cubic crystal (in units of $2\pi/A_0$)

Direction			Permutations				Magnitude
$\langle 000 \rangle$	$[000]$						0
$\langle 111 \rangle$	$[111]$	$[\bar{1}11]$	$[1\bar{1}1]$	$[11\bar{1}]$	$[\bar{1}\bar{1}1]$	$[1\bar{1}\bar{1}]$	$\sqrt{3}$
	$[\bar{1}1\bar{1}]$	$[\bar{1}\bar{1}\bar{1}]$					
$\langle 200 \rangle$	$[200]$	$[020]$	$[002]$	$[\bar{2}00]$	$[0\bar{2}0]$	$[00\bar{2}]$	2
$\langle 220 \rangle$	$[220]$	$[202]$	$[022]$	$[\bar{2}20]$	$[20\bar{2}]$	$[02\bar{2}]$	$\sqrt{8}$
	$[2\bar{2}0]$	$[\bar{2}02]$	$[0\bar{2}2]$	$[\bar{2}\bar{2}0]$	$[\bar{2}0\bar{2}]$	$[0\bar{2}\bar{2}]$	
$\langle 311 \rangle$	$[311]$	$[131]$	$[113]$	$[\bar{3}11]$	$[\bar{1}31]$	$[\bar{1}13]$	$\sqrt{11}$
	$[3\bar{1}1]$	$[1\bar{3}1]$	$[1\bar{1}3]$	$[11\bar{3}]$	$[31\bar{1}]$	$[13\bar{1}]$	
	$[\bar{3}\bar{1}1]$	$[\bar{1}\bar{3}1]$	$[\bar{1}\bar{1}3]$	$[3\bar{1}\bar{1}]$	$[1\bar{3}\bar{1}]$	$[1\bar{1}\bar{3}]$	
	$[\bar{3}1\bar{1}]$	$[\bar{1}3\bar{1}]$	$[\bar{1}1\bar{1}]$	$[\bar{3}\bar{1}\bar{1}]$	$[\bar{1}\bar{3}\bar{1}]$	$[\bar{1}\bar{1}\bar{3}]$	
$\langle 222 \rangle$	$[222]$	$[\bar{2}22]$	$[2\bar{2}2]$	$[22\bar{2}]$	$[\bar{2}\bar{2}2]$	$[2\bar{2}\bar{2}]$	$\sqrt{12}$
	$[\bar{2}2\bar{2}]$	$[\bar{2}\bar{2}\bar{2}]$					
$\langle 400 \rangle$	$[400]$	$[040]$	$[004]$	$[\bar{4}00]$	$[0\bar{4}0]$	$[00\bar{4}]$	4

It should be noted that, as q (the magnitude of $\mathbf{q}$) is a scalar, the form factors are spherically symmetric. Physically, they represent the central nuclear potential, plus the potential of the inner electron shells, and so for Si, for example, this would include the $1s^2 2s^2 2p^6$ electrons. The remaining four valence electrons, which in an isolated Si atom are found in the 3s and 3p orbitals, are the subject of the investigation; it is their energy levels and charge distributions that determine the electronic properties of the crystal and thus provide the motivation for this theoretical derivation.

The *empirical* nature of the pseudo-potential method is incorporated by adjusting the values of $V_{\mathrm{f}}(q)$ in order to achieve the closest agreement of the calculated energy levels with those measured by experimental methods. Note, therefore, that $V_{\mathrm{f}}(q)$ summarises many of the

microscopic electrostatic properties of the crystal. For example, it accounts for the nuclear charge, the inner-shell electrons, the screening provided by these electrons, and, under the auspices of the independent electron approximation described earlier, it also accounts for the electron–electron interaction experienced between the valence electrons.

As mentioned above, **q** (the difference between two reciprocal vectors) is also a reciprocal lattice vector, and in a bulk crystal, such as a face-centred cubic crystal, it takes discrete values as deduced in Chapter 1 (see Table 15.1). The pseudo-potential form factor $V_{\mathrm{f}}(q)$ is, therefore, also a discrete function, only having non-zero values for particular q. Cohen and Bergstresser [4] found that the experimentally determined band structure features of Si could be reproduced by using the values of $V_{\mathrm{f}}(q)$ given in Table 15.2.

Table 15.2 The form factors (in eV) of the common group IV semiconductor elements, converted from the original values of Cohen and Bergstresser [4]

Material	A_0 (Å)	$V_{\mathrm{f}}(\sqrt{3})$	$V_{\mathrm{f}}(\sqrt{8})$	$V_{\mathrm{f}}(\sqrt{11})$
Si	5.43	-1.43	+0.27	+0.54
Ge	5.66	-1.57	+0.07	+0.41

It should be noted that, as $V_{\mathrm{f}}(q)$ is truncated for $q > \sqrt{11}$ and as $V_{\mathrm{f}}(0)$ only has the effect of shifting the energies up or down, $V_{\mathrm{f}}(q)$ has only three non-zero values, which occur for $q = \sqrt{3}, \sqrt{8}, \sqrt{11}$. Note also that, for $q = 2$, the structure factor for the face-centred cubic crystals of interest here, which is given by equation (15.38), becomes:

$$S = 2N \cos\left(\mathbf{q}_{\bullet}\left(\frac{A_0}{8}\hat{\mathbf{i}} + \frac{A_0}{8}\hat{\mathbf{j}} + \frac{A_0}{8}\hat{\mathbf{k}}\right)\right) \tag{15.43}$$

With $\mathbf{q} = (2\pi/A_0)2\hat{\mathbf{i}}$, for example, this then gives:

$$S = 2N \cos\left(\frac{2\pi}{A_0}\frac{A_0}{8}2\hat{\mathbf{i}}_{\bullet}\hat{\mathbf{i}}\right) = 2N \cos\frac{\pi}{2} \tag{15.44}$$

which is zero, and hence the value of $V_{\mathrm{f}}(q)$ is irrelevant.

Reverting to the mathematics, then:

$$V = 2V_{\mathrm{f}}(q) \cos\left(\mathbf{G} - \mathbf{G}'\right)_{\bullet}\mathbf{T} \tag{15.45}$$

which when substituted back into equation (15.25) finally gives the complete form for the matrix elements as:

$$H_{\mathbf{G}',\mathbf{G}} = \frac{\hbar^2}{2m_0}|\mathbf{G} + \mathbf{k}|^2\delta_{\mathbf{G},\mathbf{G}'} + 2V_{\mathrm{f}}(q) \cos\left(\mathbf{G} - \mathbf{G}'\right)_{\bullet}\mathbf{T} \tag{15.46}$$

Figure 15.2 displays the results of calculations of the bulk band structure of Si by using the form factors of Table 15.2, together with the 65-element plane wave basis set in Table 15.1. The graph shows the three highest-energy valence-band levels, which (in this

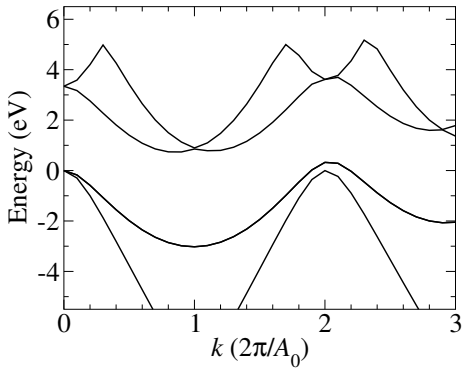


Figure 15.2: The band structure of bulk Si along the $\langle 100 \rangle$ direction, obtained by using the pseudo-potential form factors of Cohen and Bergstresser [4], adjusted to set the top of the valence band to zero

simple approach, ignoring spin–orbit coupling) are all degenerate at $k = 0$, and two of which remain degenerate across all k. In addition, the two lowest-energy conduction-band states can be clearly seen. In this first calculation, the energy levels are plotted along one of the $\langle 100 \rangle$ directions. The continuous energy *band* nature of the solutions is visible, as expected, which is in contrast to the solutions of the Schrödinger equation in heterostructures, the focus of the majority of this work so far. Furthermore, Fig. 15.2 demonstrates the periodic nature of the band structure, with equal energy levels separated by an electron wave vector equal to [200], i.e. the smallest reciprocal lattice vector along the direction of interest (see Table 15.1). Such periodicity is, of course, merely reflecting the Brillouin zone symmetry structure of the crystal [3] and the edge of the *first* Brillouin zone is at exactly half of the reciprocal lattice vector in question, i.e. $k = 2\pi/A_0$. Note that there is a slight discrepancy in the energy levels, e.g. exact degeneracy is not reproduced at the top of the valence band at $k = 2 \times 2\pi/A_0$; this is simply a computational deficiency and, as the vast majority of calculations are confined to the first Brillouin zone, is not an issue here.

Of course, there is a periodicity in the energy bands along other crystal directions, e.g. the edge of the Brillouin zone along the $\langle 111 \rangle$ direction is at $k = \sqrt{3}/2 \times 2\pi/A_0$. The data from these two principal axes of symmetry are often gathered together to produce a standard form for displaying the band structure; Fig. 15.3 illustrates this for the two most common group IV elemental semiconductors, i.e. Si and Ge. Again, the form factors given in Table 15.2 were employed, together with the 65-element basis set of Table 15.1. The indirect nature of both materials can be seen from the bandgap minimum for Si occurring towards the X point within the Brillouin zone, while for Ge it occurs towards L. The choice of the basis set, i.e. plane waves, has been discussed, and the *number* 65 can be justified by the data presented in Table 15.3. This table displays the results of calculations of the energy of the lowest conduction band in Si at two electron wave vectors $\mathbf{k}$. The first is for the Γ-point ($\mathbf{k} = 0$), and the second is near the X minima ($\mathbf{k} = 0, 0, 0.8$). In addition, the table shows the discretisation of the face-centred reciprocal lattice vectors, plus the largest vector included in the set.

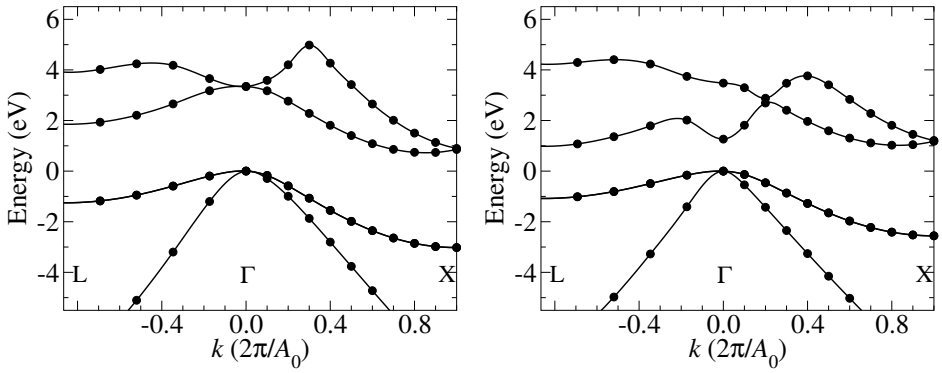


Figure 15.3: The band structures of bulk Si (left) and Ge (right) obtained by using the pseudo-potential form factors of Cohen and Bergstresser [4], adjusted to set the top of the valence band to zero

Table 15.3 The energies of the lowest conduction band of Si at Γ and [0,0,0.8] (near the X minima) as a function of the number of plane waves in the basis set

Number of plane waves	Largest **G** $(2\pi/A_0)$	E_Γ (eV)	E_X (eV)
15	⟨200⟩	3.212	0.926
27	⟨220⟩	3.453	0.982
51	⟨311⟩	3.552	1.085
59	⟨222⟩	3.389	0.827
65	⟨400⟩	3.344	0.741
89	⟨331⟩	3.382	0.781
113	⟨420⟩	3.412	0.825
137	⟨422⟩	3.409	0.821

The convergence of the band structure is perhaps most visible through a plot of the above data, as in Fig. 15.4. The data corresponding to a 65-element basis set is just at the bottom of the shoulder, i.e. at the beginning of the plateau. It is this accuracy that is generally chosen for simple bulk calculations, such as in Fig. 15.3.

In summary, the *empirical* pseudo-potential method is a technique for inputting the experimentally determined features into a theoretical model of the electronic structure of the crystal. At this level, the features that are reproduced are the fundamental bandstructure properties, such as the band gap at the centre of the Brillouin Zone, Γ, and at indirect gaps such as the X valley in Si.

15.3 Spin–orbit coupling

Speaking simplistically, the four valence electrons of a Si (or Ge) atom, three of which occupy p-states and one an s-state, produce four equivalent covalent bonds (sp^3 hybridisation)

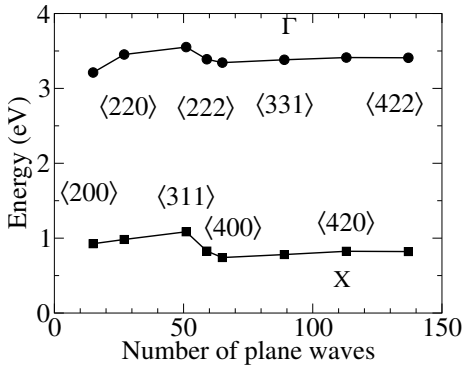


Figure 15.4: The energies of the lowest conduction band of Si at Γ and [0,0,0.8] (near the X minima) as a function of the number of plane waves in the basis set

directed towards the corners of a tetrahedron, upon crystallisation. Four neighbouring atoms each supply a single electron, and hence any one atom completes its outer shell of eight electrons by forming four doubly occupied bonds. The same is also true for a compound semiconductor, but in this case one species, the cation (e.g. Ga), supplies three electrons and the anion (e.g. As) supplies five.

Within an infinite crystal, as has already been shown with the pseudo-potential calculations in the previous section, these bonds form *bands*. Figure 15.5 illustrates the three uppermost of these *valence bands* for Si and Ge. The lowest-energy valence band is far off scale to the bottom of the figure and represents a well-localised state around the atomic cores, and has an s-type nature.

The other three bands, in the calculations performed so far, are degenerate for stationary electrons (holes). These bands resemble the p-type atomic orbitals and have an orbital angular momentum, $L = 1$ (see, for example, [5], p. 41). This couples with the spin angular momentum, $S = \frac{1}{2}$, to give $J = L + S = \frac{3}{2}$ and $J = L - S = \frac{1}{2}$, thus lifting the degeneracy. This is analogous to LS coupling in a low-atomic-number atom (see any atomic physics book for an introduction, for example [6], pp. 428 and 441).

In III–V compounds, it has been found that the quadruplet $J = \frac{3}{2}$ is higher in energy than the $J = \frac{1}{2}$ doublet. Note that the degeneracy, whether twofold or fourfold, arises from the number of possible values that the projection J_z of J can take. As the total energy remains the same, the change in the energy of the quadruplet $J = \frac{3}{2}$ must be half that of the $J = \frac{1}{2}$ doublet.

The quadruplet is degenerate in bulk systems at the zone centre, but for an increasing electron wave vector it splits into two separate spin-degenerate bands. The lower-energy band corresponds to the $m_J = \frac{3}{2}$ projection and, because of the effective mass interpretation of the E–$\mathbf{k}$ curves, it has a higher effective mass than the higher-energy $m_J = \frac{1}{2}$ band. Hence, the two $J = \frac{3}{2}$ bands are labelled as the heavy-hole (HH, $m_J = \frac{3}{2}$) and the light-hole (LH, $m_J = \frac{1}{2}$) bands, as shown in Fig. 15.5.

The spin–orbit interaction can be accounted for within the empirical pseudo-potential

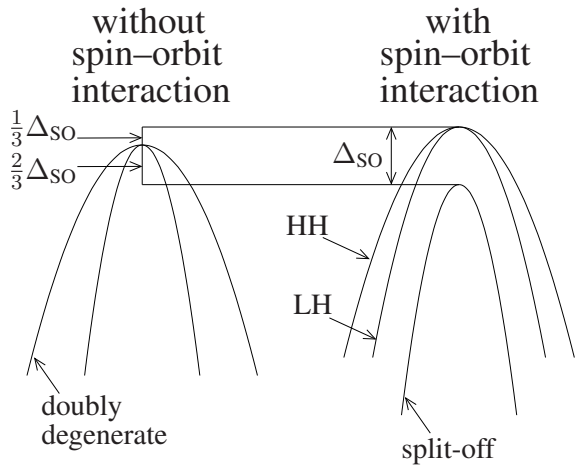


Figure 15.5: The effect of the spin–orbital interaction on the valence-band structure

approach [7, 8] and amounts to an addition term in the potential V as originally defined in equation (15.26), i.e. the crystal potential V_c is replaced with $V_c + V_{SO}$.

Inclusion of the spin–orbit interaction is essential if detailed calculations of valence-band structure are to be performed, and this will be returned to after the elemental calculation has been generalised to compound semiconductors.

15.4 Compound semiconductors

The generalisation of the calculation to compound semiconductors manifests itself merely as a change in the atomic basis.

Recalling equation (15.31):

$$V = \frac{1}{\Omega} \sum_{\mathbf{r}_a} \mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{r}_a} \int V_a(\mathbf{r}) \mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{r}} \, \mathrm{d}\tau \tag{15.47}$$

then again, the structure factor can be written as:

$$S = \sum_{\mathbf{r}_a} \mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{r}_a} = \sum_{\mathbf{R}} \sum_{\mathbf{t}} \mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet(\mathbf{R}+\mathbf{t})} \tag{15.48}$$

giving, as before:

$$S = N \sum_{\mathbf{t}} \mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{t}} \tag{15.49}$$

However, for a compound semiconductor, the sum over $\mathbf{t}$ is over two dissimilar atoms (the cation is at $-\mathbf{T}$ and the anion is at $+\mathbf{T}$, where $\mathbf{T} = \frac{A_0}{8}(\hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}})$) and therefore it cannot be made independent of the form of the atomic potential $V_a(\mathbf{r})$. Hence, the structure factor S

must be substituted back into equation (15.47) at this point, thus giving:

$$V = \frac{N}{\Omega} \sum_{\mathbf{t}} \mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{t}} \int V_a(\mathbf{r}) \mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{r}} \, \mathrm{d}\tau \tag{15.50}$$

Summing over the two basis positions, $\mathbf{T}^{\text{cat}} = -\mathbf{T}$, and $\mathbf{T}^{\text{an}} = +\mathbf{T}$, then:

$$V = \frac{N}{\Omega} \left[\mathrm{e}^{-i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{T}} \int V_a^{\text{cat}}(\mathbf{r}) \mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{r}} \, \mathrm{d}\tau + \mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{T}} \int V_a^{\text{an}}(\mathbf{r}) \mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{r}} \, \mathrm{d}\tau \right] \tag{15.51}$$

Recalling that $N/\Omega = 1/\Omega_{\text{c}}$ and again writing $\mathbf{q} = \mathbf{G}' - \mathbf{G}$ leads to:

$$V = \mathrm{e}^{-i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{T}} \frac{1}{\Omega_{\text{c}}} \int V_a^{\text{cat}}(\mathbf{r}) \mathrm{e}^{-i\mathbf{q}\bullet\mathbf{r}} \, \mathrm{d}\tau + \mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{T}} \frac{1}{\Omega_{\text{c}}} \int V_a^{\text{an}}(\mathbf{r}) \mathrm{e}^{-i\mathbf{q}\bullet\mathbf{r}} \, \mathrm{d}\tau \tag{15.52}$$

Now the definition of the pseudo-potential form factor (from equation (15.42)) is as follows:

$$V_{\text{f}}(q) = \frac{1}{\Omega_{\text{c}}} \int V_a(\mathbf{r}) \mathrm{e}^{-i\mathbf{q}\bullet\mathbf{r}} \, \mathrm{d}\tau \tag{15.53}$$

$$\therefore V = V_{\text{f}}^{\text{cat}}(q) \mathrm{e}^{-i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{T}} + V_{\text{f}}^{\text{an}}(q) \mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{T}} \tag{15.54}$$

where the form factors associated with the cation and anion atomic potentials, $V_a^{\text{cat}}(\mathbf{r})$ and $V_a^{\text{an}}(\mathbf{r})$, have been labelled as $V_{\text{f}}^{\text{cat}}(q)$ and $V_{\text{f}}^{\text{an}}(q)$, respectively. This equation can be manipulated further to give:

$$V = \left[V_{\text{f}}^{\text{an}}(q) + V_{\text{f}}^{\text{cat}}(q)\right] \cos\left(\mathbf{G} - \mathbf{G}'\right)\bullet\mathbf{T} + i\left[V_{\text{f}}^{\text{an}}(q) - V_{\text{f}}^{\text{cat}}(q)\right] \sin\left(\mathbf{G} - \mathbf{G}'\right)\bullet\mathbf{T} \tag{15.55}$$

A quick check on the validity of this extension is to put the cation potential $V_{\text{f}}^{\text{cat}}(q)$ and the anion potential $V_{\text{f}}^{\text{an}}(q)$ equal to the same potential $V_{\text{f}}(q)$, thus reproducing the potential term of the elemental calculation, as in equation (15.45).

Often the 'sum' $[V_{\text{f}}^{\text{an}}(q) + V_{\text{f}}^{\text{cat}}(q)]$ and 'difference' $[V_{\text{f}}^{\text{an}}(q) - V_{\text{f}}^{\text{cat}}(q)]$ potentials are relabelled as the symmetric and anti-symmetric form factors, respectively, i.e. $V_{\text{f}}^{\text{S}}(q)$ and $V_{\text{f}}^{\text{A}}(q)$. This gives:

$$V = V_{\text{f}}^{\text{S}}(q) \cos\left(\mathbf{G} - \mathbf{G}'\right)\bullet\mathbf{T} + iV_{\text{f}}^{\text{A}}(q) \sin\left(\mathbf{G} - \mathbf{G}'\right)\bullet\mathbf{T} \tag{15.56}$$

Substituting for the potential V into equation (15.25) finally gives all of the elements of the Hamiltonian matrix as:

$$H_{\mathbf{G}',\mathbf{G}} = \frac{\hbar^2}{2m_0} |\mathbf{G} + \mathbf{k}|^2 \delta_{\mathbf{G},\mathbf{G}'} + V_{\text{f}}^{\text{S}}(q) \cos\left(\mathbf{G} - \mathbf{G}'\right)\bullet\mathbf{T} + iV_{\text{f}}^{\text{A}}(q) \sin\left(\mathbf{G} - \mathbf{G}'\right)\bullet\mathbf{T} \tag{15.57}$$

Table 15.4 gives examples of the symmetric ($V_{\text{f}}^{\text{S}}(q)$) and anti-symmetric ($V_{\text{f}}^{\text{A}}(q)$) form factors as deduced *empirically* by Cohen and Bergstresser [4]. Note that the value of the symmetric form factor of $q = 2$, i.e. $V_{\text{f}}^{\text{S}}(2)$, is again unnecessary because of a zero structure factor, as in the elemental (group IV) calculation. In the anti-symmetric case, however, the structure factor corresponding to $q = \sqrt{8}$ is zero, thus allowing this form factor to remain undefined.

Table 15.4 The form factors (eV) of a selection of III–V compound semiconductors, converted from the original values of Cohen and Bergstresser [4]

Material	$V_{\rm f}^{\rm S}(q)$ $\sqrt{3}$	$\sqrt{8}$	$\sqrt{11}$	$V_{\rm f}^{\rm A}(q)$ $\sqrt{3}$	2	$\sqrt{11}$
GaAs	−3.13	+0.14	+0.82	+0.95	+0.68	+0.14
InAs	−2.99	0.00	+0.68	+1.09	+0.68	+0.41
GaP	−2.99	+0.41	+0.95	+1.63	+0.95	+0.27
InP	−3.13	+0.14	+0.82	+0.95	+0.68	+0.14

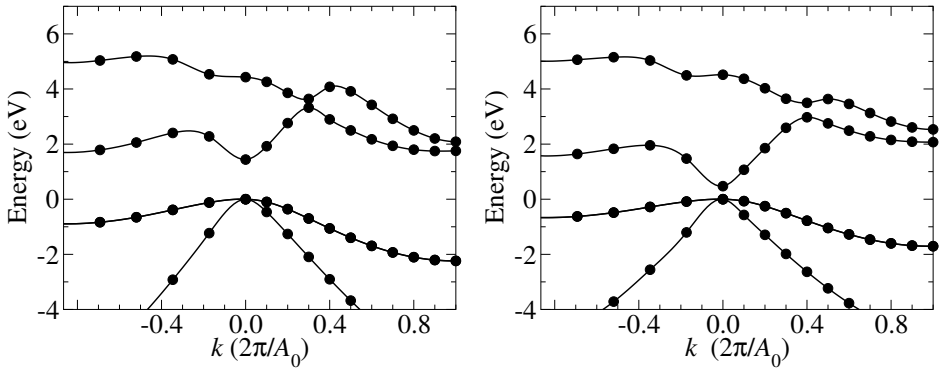


Figure 15.6: The band structures of bulk GaAs (left) and InAs (right) obtained by using the pseudo-potential form factors of Cohen and Bergstresser [4], shifted to set the top of the valence band to zero

Fig. 15.6 displays the results of calculations of the band structures of bulk GaAs and InAs. In contrast to Si and Ge, the minimum band gaps are direct, and thus an electron and hole can recombine without recourse to a momentum change, resulting in efficient light emission. Hence, the use of both of these materials as part of the quaternary compound $\mathrm{In}_{1-x}\mathrm{Ga}_x\mathrm{As}_y\mathrm{P}_{1-y}$ [9] in light-emitting diodes [10].

In recent years, a new material, GaN (and its alloys with In), has come to prominence with the development of room-temperature blue-light-emitting diodes [11] and continuous-wave lasers [12, 13]. GaN is particularly interesting from the viewpoint of electronic bandstructure calculations in that it does not take the face-centred cubic zinc blende or diamond structures employed so far in this work, rather the natural crystal structure of GaN, being the hexagonal close-packed wurtzite structure (see, for example, [14], p. 47).

Such a change in crystal symmetry does not represent a problem for the analysis derived above, as it is only necessary to change the basis vectors, the reciprocal lattice vectors and, of course, employ the appropriate pseudo-potential form factors (for such a treatise, see, for example, [15]). Many examples will appear in the following, which illustrate pseudo-potential calculations of systems of symmetry other than zinc blende, e.g. those involving

isolated impurities, superlattices and quantum wires.

15.5 Charge densities

In addition to yielding the eigenvalues (energy levels), the diagonalisation of equations (15.46) and (15.57) also gives the corresponding eigenvectors, i.e. the expansion coefficients $a_{n,\mathbf{k}}$ for each of the basis vectors $\mathbf{G}$. It is straightforward to generate the wave function from these eigenvectors by using equation (15.9), i.e.

$$\psi_{n,\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} a_{n,\mathbf{k}}(\mathbf{G}) \mathrm{e}^{i(\mathbf{G}+\mathbf{k})\bullet\mathbf{r}} \tag{15.58}$$

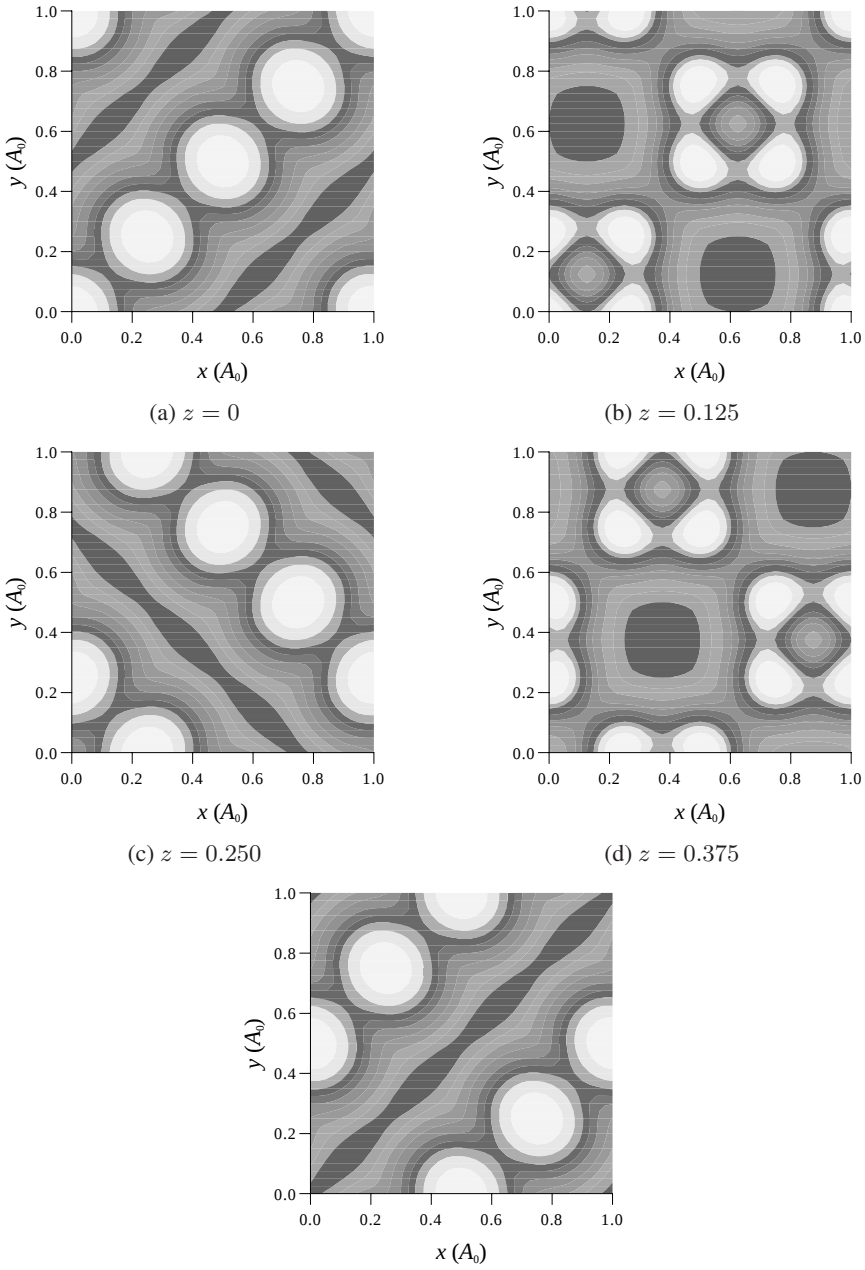
Utilising the probability interpretation of the wave function, i.e. that the probability of finding a particle at a point is proportional to $\psi^*\psi$, the *charge density* is given by:

$$\rho = \psi^*_{n,\mathbf{k}}(\mathbf{r})\psi_{n,\mathbf{k}}(\mathbf{r}) \tag{15.59}$$

and can be easily evaluated.

Figure 15.7 illustrates the results of summing the charge density over all of the four bulk valence-band states, across a series of x–y planes through a single face-centred cube of Si. Given the coordinate system employed thus far, the Bravais lattice points are, of course, at the corners and on the faces of the cube of side A_0, with a pair of atoms at $(-\frac{1}{8}, -\frac{1}{8}, -\frac{1}{8})$ and $(+\frac{1}{8}, +\frac{1}{8}, +\frac{1}{8})$, respectively. Therefore, the first ($z = 0$), third ($z = 0.250A_0$) and fifth ($z = 0.5A_0$) planes are cross-sections through the bond centres. The symmetry of the sp^3-hybridised bonds can be seen most clearly across the planes intersecting the centre of atoms, namely the second ($z = 0.125A_0$) and fourth ($z = 0.375A_0$) planes. In addition, this figure illustrates the equality of the atoms in this, the diamond structure, where the symmetry between the planes containing the atoms is merely translational.

In contrast to this, Fig. 15.8 displays the corresponding charge density plot, again summed over all four valence-band states, but this time for the zinc blende structure of GaAs. Again, there is a symmetry between the planes intersecting the bond centres, i.e. the first ($z = 0$), third ($z = 0.250A_0$) and fifth ($z = 0.5A_0$). However, the two planes intersecting the centres of the atoms are quite dissimilar. In particular, the second ($z = 0.125A_0$) plane contains the anions, in this case As, and is quite different from the fourth ($z = 0.375A_0$) plane which contains the cations, i.e. Ga. The fourth plane displays an increase in the charge density (represented by the lighter colour) around the Ga atoms, while the second plane displays a decrease in the charge density around the As atoms, in comparison to the Si structure of Fig. 15.7. This is a reflection of the increased *ionicity* in the Ga–As bond, in comparison with the completely covalent bond in bulk Si. Instead of having an equal share of the two valence electrons that sit between the atomic cores and form the bond, as occurs in the diamond structure, the Ga atom takes a larger share, thus becoming a partially negatively charged centre (i.e. a cation); the As atom therefore has a smaller share, thus becoming a partially positively charged centre (i.e. an anion).

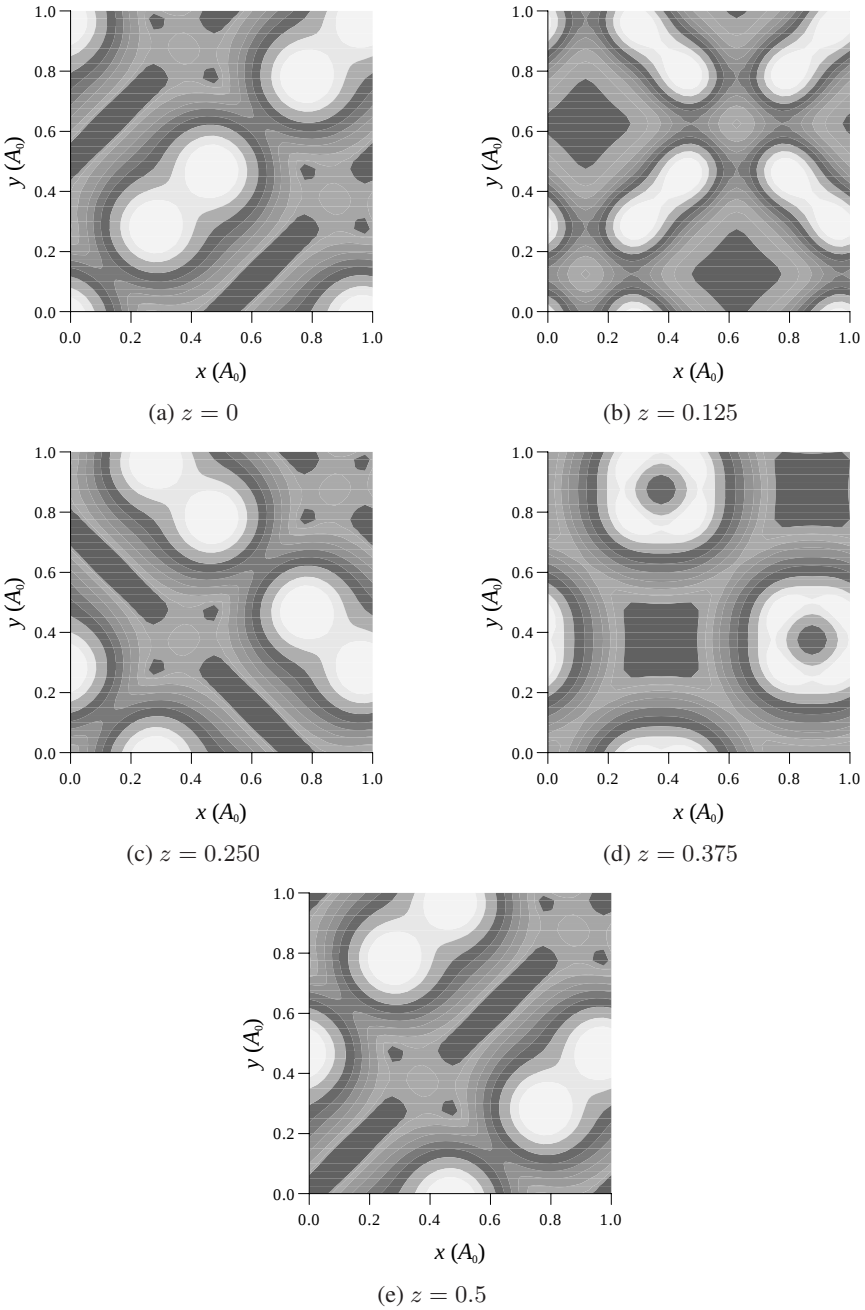


(a) $z = 0$ (b) $z = 0.125$

(c) $z = 0.250$ (d) $z = 0.375$

(e) $z = 0.5$

Figure 15.7: Total valence-band charge density for Si across the x–y plane of a single face-centred cubic unit cell, for a variety of z values (given in A_0); note that the lighter the colour, the higher the charge density. The Bravais lattice points are at the corners and centre of the $z = 0$ plane and on the edges of the $z = 0.5A_0$ plane



(a) $z = 0$ (b) $z = 0.125$

(c) $z = 0.250$ (d) $z = 0.375$

(e) $z = 0.5$

Figure 15.8: Total valence-band charge density for GaAs across the x–y plane of a single face-centred cubic unit cell, for a variety of z values (given in A_0); note that the lighter the colour, the higher the charge density. The Bravais lattice points are at the corners and centre of the $z = 0$ plane and on the edges of the $z = 0.5A_0$ plane

15.6 Calculating the effective mass

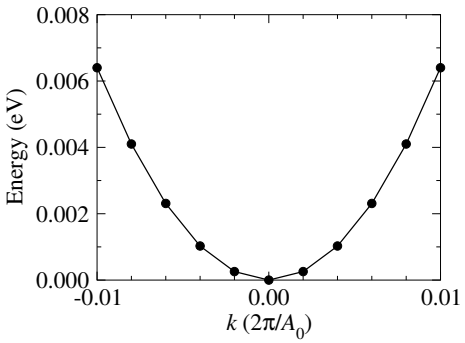


Figure 15.9: Empirical pseudo-potential calculations of the band structure of bulk GaAs around the centre (Γ) of the Brillouin zone (solid circles), together with a parabolic fit to the data (line)

Empirical pseudo-potential calculations can be used to calculate material parameters other than just the main energy gaps. Figure 15.9 displays the results of calculations of the lowest conduction band of GaAs at points along one of the $\langle 100 \rangle$ directions around the centre of the Brillouin zone (the minimum has been adjusted to zero). In addition to the data points, a parabolic fit $E = Ak^2$ has also been included, suggested by the definition of the effective mass, i.e.

$$E = \frac{\hbar^2 k^2}{2m^*} \tag{15.60}$$

From this, the effective mass m^* of the electrons around the Γ conduction-band minimum can be calculated, in this case giving $m^* = 0.074m_0$. Given the crudity of the form factors employed, this compares well with the accepted value of $0.067m_0$ [16]. This implies that form factors that have been deduced empirically by comparing calculated energy gaps with those measured experimentally can also give good information as to the effective masses. This approach has been used successfully in other materials, and in one case has produced additional insight into the behaviour of excitons in quantum wells [17]. More sophisticated potentials will be introduced later, which, in addition to allowing for more accurate determination of effective masses, will allow for the calculation of deformation potentials in strained crystals.

15.7 Alloys

Mixing two binary (or indeed elemental) semiconductors together to form an alloy is a very common and often used technique for producing a whole new range of materials whose fundamental properties, e.g. the band gap and the Γ valley effective mass, can be tuned by adjusting the proportions of the constituents [16]. The simplest semiconductor alloy would

be that formed from the two common elemental semiconductors Si and Ge [18], but perhaps the most common are the alloys of GaAs, such as $Ga_{1-x}Al_xAs$, a ternary alloy formed from GaAs and AlAs.

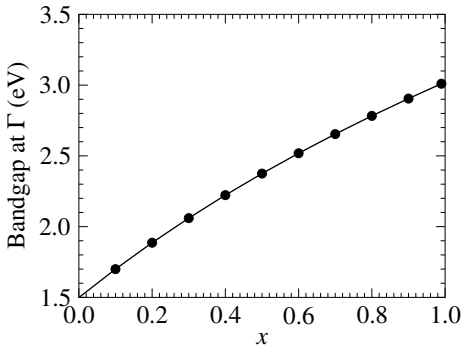


Figure 15.10: The zone-centre band gap of a $Ga_{1-x}Al_xAs$ alloy calculated by linearly interpreting the pseudo-potential form factors

Generally, such alloys are assumed to have properties that vary linearly between the two constituents, a procedure that is known as the *virtual crystal approximation*, thus giving material parameters that are simply taken in proportion from their components. For example, the band gap of $Ga_{1-x}Al_xAs$ is given by:

$$E_g^{Ga_{1-x}Al_xAs} = (1-x)E_g^{GaAs} + xE_g^{AlAs} \tag{15.61}$$

However, in some instances such a linear interpolation is only approximate and the actual dependence is more complicated, e.g. it is quite surprising that there is still some disagreement as to the particular form of the band gap of $Ga_{1-x}Al_xAs$, as highlighted by the range of measured values quoted by Adachi ([16], p. 146). A nonlinear dependence is often written as:

$$E_g^{Ga_{1-x}Al_xAs} = (1-x)E_g^{GaAs} + xE_g^{AlAs} + C_{GaAl}x(x-1) \tag{15.62}$$

where $C_{GaAl}x(x-1)$ is known as the *bowing parameter*.

Such detailed reviews of the properties of alloys lie outside the scope of the present work; suffice it to say here that the particulars are well documented (see, for example, [18, 16, 19]) and that in this work it is sufficient to know that equation (15.61) can be generalised to represent any material parameter, e.g. the lattice constant:

$$A_0^{Ga_{1-x}Al_xAs} = (1-x)A_0^{GaAs} + xA_0^{AlAs} \tag{15.63}$$

or the pseudo-potential form factors:

$$V_f(q)^{Ga_{1-x}Al_xAs} = (1-x)V_f(q)^{GaAs} + xV_f(q)^{AlAs} \tag{15.64}$$

Fig. 15.10 displays the results of a simple demonstration calculation showing the band gap at the centre of the Brillouin zone as a function of the proportion of Al in the $Ga_{1-x}Al_xAs$ alloy; note that a slight degree of bowing *is* evident here.

The reader may be interested to know that the microscopic properties of alloys are themselves a thriving research topic, one aspect of which is whether the constituent atoms, for example, Si and Ge or Ga and Al, are arranged randomly or otherwise (see, for example, [20, 21, 22]).

15.8 Atomic form factors

It is interesting to decompose the symmetric and anti-symmetric form factors to yield the individual *atomic* form factors, which are given simply by:

$$V_{\mathrm{f}}^{\mathrm{an}}(q) = \frac{1}{2}\left[V_{\mathrm{f}}^{\mathrm{S}}(q) + V_{\mathrm{f}}^{\mathrm{A}}(q)\right] \tag{15.65}$$

$$V_{\mathrm{f}}^{\mathrm{cat}}(q) = \frac{1}{2}\left[V_{\mathrm{f}}^{\mathrm{S}}(q) - V_{\mathrm{f}}^{\mathrm{A}}(q)\right] \tag{15.66}$$

The term 'atomic' is introduced here to specify unambiguously that these are pseudo-potential form factors that can be identified with an individual atom, unlike the symmetric, $V_{\mathrm{f}}^{\mathrm{S}}(q)$, and anti-symmetric, $V_{\mathrm{f}}^{\mathrm{A}}(q)$, form factors which are identified with a cation–anion basis.

Table 15.5 The atomic form factors (eV) of a selection of III–V compound semiconductors, decomposed from their original symmetric and anti-symmetric values as in Table 15.4

Cation	Anion	$V_{\mathrm{f}}^{\mathrm{cat}}(q)$		$V_{\mathrm{f}}^{\mathrm{an}}(q)$	
		$\sqrt{3}$	$\sqrt{11}$	$\sqrt{3}$	$\sqrt{11}$
Ga	As	−2.04	+0.34	−1.09	+0.48
In	As	−2.04	+0.14	−0.95	+0.54
Ga	P	−2.31	+0.34	−0.68	+0.61
In	P	−2.04	+0.34	−1.09	+0.48

Table 15.5 gives the atomic form factors for the cations (Ga, In) and anions (As, P) calculated in this manner. Note at this point that it is not possible to deduce the atomic form factors $V_{\mathrm{f}}^{\mathrm{an}}(q)$ and $V_{\mathrm{f}}^{\mathrm{cat}}(q)$ at $q = 2$ and $\sqrt{8}$ as this would rely on *both* the symmetric ($V_{\mathrm{f}}^{\mathrm{S}}(q)$) and anti-symmetric ($V_{\mathrm{f}}^{\mathrm{A}}(q)$) form factors being known at these points. There are quite some differences between the potentials of the same species in different materials, e.g. the As associated with GaAs, compared with the As associated with InAs. This appears to be a general result, with the physical interpretation being that the *core* potential, i.e. nucleus plus inner electron shells, of the As atom is indeed different in both materials, with this difference arising from the crystal environment. Such a result is disappointing, as a single well-defined atomic potential suitable for all compounds would be a very powerful parameter indeed, allowing the properties of new materials to be predicted without recourse to empirical deductions of the form factors.

15.9 Generalisation to a large basis

Thus far, the background to the empirical pseudo-potential method has been explored and applied to the calculation of the electronic band structure of bulk elemental and compound semiconductors. It is convenient now to generalise this method to allow for the calculation of the electronic properties of systems of different symmetries, in particular quantum wells, wires and dots. The application of this theory will be covered in the following chapter.

Consider now the first, and hence most general, expression for the crystal potential (equation (15.31)), i.e.

$$V = \frac{1}{\Omega}\sum_{\mathbf{r}_a} \mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{r}_a}\int V_a(\mathbf{r})\mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{r}}\,\mathrm{d}\tau \tag{15.67}$$

Again, writing $\mathbf{q} = \mathbf{G}' - \mathbf{G}$:

$$V = \frac{\Omega_\mathrm{c}}{\Omega}\sum_{\mathbf{r}_a} \mathrm{e}^{-i\mathbf{q}\bullet\mathbf{r}_a}\frac{1}{\Omega_\mathrm{c}}\int V_a(\mathbf{r})\mathrm{e}^{-iq\bullet\mathbf{r}}\,\mathrm{d}\tau \tag{15.68}$$

where Ω_c is again the volume of the primitive cell, which in the case of a face-centred cubic crystal is $A_0^3/4$. Now clearly the normalised integral of the atomic potential $V_a(\mathbf{r})$ is still the pseudo-potential form factor, although now it is acknowledged that the generalisation may allow for many atom types at as yet unspecified positions. Hence, it is important to write:

$$\frac{1}{\Omega_\mathrm{c}}\int V_a(\mathbf{r})\mathrm{e}^{-iq\bullet\mathbf{r}}\,\mathrm{d}\tau \quad \text{as} \quad V_\mathrm{f}^{\mathbf{r}_a}(q) \tag{15.69}$$

implying the empirical pseudo-potential form factor of the atom at position $\mathbf{r}_a$. Therefore:

$$V = \frac{\Omega_\mathrm{c}}{\Omega}\sum_{\mathbf{r}_a} \mathrm{e}^{-i\mathbf{q}\bullet\mathbf{r}_a}\ V_\mathrm{f}^{\mathbf{r}_a}(q) \tag{15.70}$$

As with bulk semiconductors, the atomic positions can always be written as a sum of a Bravais lattice vector and a basis vector, i.e.

$$\mathbf{r}_a = \mathbf{R} + \mathbf{t} \tag{15.71}$$

which, as before, replaces the summation over the atomic positions $\mathbf{r}_a$ with two summations, one over $\mathbf{R}$ and one over $\mathbf{t}$, thus giving the following:

$$V = \frac{\Omega_\mathrm{c}}{\Omega}\sum_{\mathbf{R}} \mathrm{e}^{-i\mathbf{q}\bullet\mathbf{R}}\sum_{\mathbf{t}} \mathrm{e}^{-i\mathbf{q}\bullet\mathbf{t}}\ V_\mathrm{f}^{\mathbf{r}_a}(q) \tag{15.72}$$

Clearly, whatever the symmetry, the expansion set of plane waves will always have the periodicity of the system, so, just as before, the scalar product of a reciprocal lattice vector, $\mathbf{q} = \mathbf{G}' - \mathbf{G}$, with a Bravais lattice vector, $\mathbf{R}$, will be equal to an integral multiple of 2π, and hence $\mathrm{e}^{-i\mathbf{q}\bullet\mathbf{R}} = 1$. If there are N of these new generalised bases in the total volume of the crystal, then:

$$V = \frac{\Omega_\mathrm{c}N}{\Omega}\sum_{\mathbf{t}} \mathrm{e}^{-i\mathbf{q}\bullet\mathbf{t}}\ V_\mathrm{f}^{\mathbf{r}_a}(q) \tag{15.73}$$

Table 15.6 The form factors (eV) of a selection of III–V compound semiconductors, converted from the original values of Cohen and Bergstresser [4], where the symmetric form factor at $q = 2$ and the anti-symmetric at $q = \sqrt{8}$ have been deduced by linear interpolation from the two adjacent values

	$V_{\rm f}^{\rm S}(q)$				$V_{\rm f}^{\rm A}(q)$			
	$\sqrt{3}$	2	$\sqrt{8}$	$\sqrt{11}$	$\sqrt{3}$	2	$\sqrt{8}$	$\sqrt{11}$
GaAs	−3.13	*−2.33*	+0.14	+0.82	+0.95	+0.68	*+0.34*	+0.14
InAs	−2.99	*−2.26*	0.00	+0.68	+1.09	+0.68	*+0.51*	+0.41
GaP	−2.99	*−2.16*	+0.41	+0.95	+1.63	+0.95	*+0.52*	+0.27
InP	−3.13	*−2.33*	+0.14	+0.82	+0.95	+0.68	*+0.34*	+0.14

Now the total volume of the crystal, Ω, divided by the number of the new general Bravais lattice points, N, is equal to the volume occupied by one of the bases, $\Omega_{\rm b}$ (say), and therefore:

$$V = \frac{\Omega_{\rm c}}{\Omega_{\rm b}} \sum_{\mathbf{t}} {\rm e}^{-i\mathbf{q}\bullet\mathbf{t}} \;\; V_{\rm f}^{\mathbf{r}_a}(q) \tag{15.74}$$

The diamond, zinc blende and wurtzite crystals discussed so far all have a two-atom basis, and hence $\Omega_{\rm c}$ (the volume of the primitive cell) has been the volume occupied by two atoms. In this more general basis with N_a atoms (say), the volume occupied by its primitive cell would be:

$$\Omega_{\rm b} = \frac{N_a}{2}\Omega_{\rm c} \tag{15.75}$$

Therefore, the final expression for the crystal potential follows as:

$$V = \frac{2}{N_a} \sum_{\mathbf{t}} {\rm e}^{-i\mathbf{q}\bullet\mathbf{t}} \;\; V_{\rm f}^{\mathbf{r}_a}(q) \tag{15.76}$$

and the full expression for the Hamiltonian matrix elements is:

$$H_{\mathbf{G}',\mathbf{G}} = \frac{\hbar^2}{2m_0}|\mathbf{G}+\mathbf{k}|^2\delta_{\mathbf{G},\mathbf{G}'} + \frac{2}{N_a} \sum_{\mathbf{t}} {\rm e}^{-i\mathbf{q}\bullet\mathbf{t}} \;\; V_{\rm f}^{\mathbf{r}_a}(q) \tag{15.77}$$

At this point, the final equation appears to be merely a re-expression of the Hamiltonian matrix elements of the elemental and compound bulk semiconductors, and it can indeed be shown to reproduce those expressions. For example, consider an ($N_a = 2$)-atom basis, with a cation at $-\mathbf{T} = -\frac{A_0}{8}(\hat{\mathbf{i}}+\hat{\mathbf{j}}+\hat{\mathbf{k}})$ and an anion at $+\mathbf{T}$. Then the potential term in equation (15.77) gives:

$$V = \frac{2}{2}\left[{\rm e}^{i\mathbf{q}\bullet\mathbf{T}} \;\; V_{\rm f}^{\rm cat}(q) + {\rm e}^{-i\mathbf{q}\bullet\mathbf{T}} \;\; V_{\rm f}^{\rm an}(q)\right] \tag{15.78}$$

which, recalling that $\mathbf{q} = \mathbf{G}' - \mathbf{G}$, is equivalent to equation (15.54).

However, the above is much more than just a generalised form for bulk semiconductor calculations; by thoughtful choice of the atomic basis and the primitive cell, equation (15.77) can be used to calculate the electronic structure of heterostructures of all dimensions,

Table 15.7 The atomic form factors (eV) of a selection of III–V compound semiconductors, decomposed from their original symmetric and anti-symmetric values and now including the interpolated potentials for $q = 2$ and $\sqrt{8}$ (as indicated by italics)

	$V_{\rm f}^{\rm cat}(q)$				$V_{\rm f}^{\rm an}(q)$			
	$\sqrt{3}$	2	$\sqrt{8}$	$\sqrt{11}$	$\sqrt{3}$	2	$\sqrt{8}$	$\sqrt{11}$
GaAs	−2.04	*−1.51*	*−0.10*	+0.34	−1.09	*−0.83*	*+0.24*	+0.48
InAs	−2.04	*−1.47*	*−0.26*	+0.14	−0.95	*−0.79*	*+0.26*	+0.55
GaP	−2.31	*−1.56*	*−0.06*	+0.34	−0.68	*−0.61*	*+0.47*	+0.61
InP	−2.04	*−1.51*	*−0.10*	+0.34	−1.09	*−0.83*	*+0.24*	+0.48

i.e. quantum wells, wires and dots. Such calculations are often referred to as *large-cell* calculations; however, the computational method of summing over atoms in a more extensive basis suggests that the term *large-basis* calculations is more appropriate. The promise of this generalisation will be explored fully in Chapter 16; however, for now, it is worthwhile pursuing these ideas for bulk systems.

In particular, the brief digression into *atomic form factors* was in anticipation of this formalism—this large-basis method relies on pseudo-potential form factors, which are associated with individual atoms, and hence, in order to perform calculations of bulk compound semiconductors, the symmetric and anti-symmetric form factors listed in Table 15.4 need to be decomposed (as in Section 15.8) into their atomic components. In addition, however, the 'structure factor' associated with these atomic potentials is now of the form $\mathrm{e}^{-i\mathbf{q}\bullet\mathbf{t}}$, which is never zero, and hence the *atomic* form factors in this method need to be specified at all values of **q** that can arise from the basis set of reciprocal lattice vectors **G**.

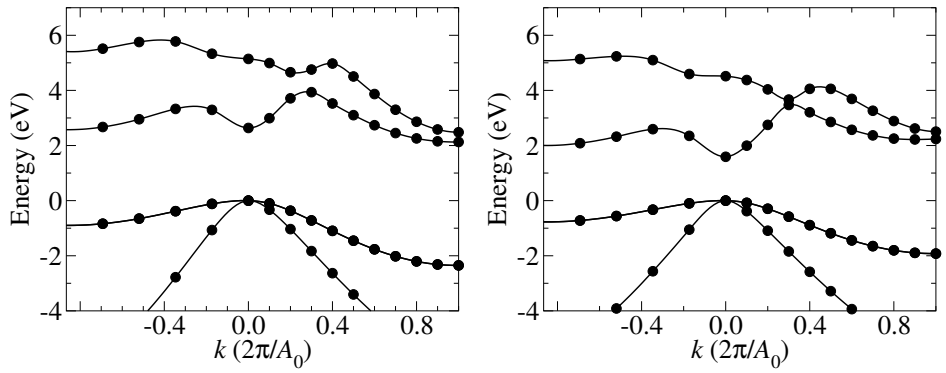


Figure 15.11: The band structures of bulk GaP (left) and InP (right) obtained by using the interpolated atomic form factors given in Table 15.7, shifted to set the top of the valence band to zero

For a compound semiconductor, this implies that it is necessary to know the atomic form factors at $q = 2$ and $\sqrt{8}$. For the example III–V materials given earlier in Table 15.4, this can

be achieved as a zeroth-order approximation simply by linearly interpolating between the existing symmetric and anti-symmetric potentials (as in Table 15.6), and then decomposing the resulting form factors to give the data presented in Table 15.7.

Figure 15.11 shows the band structures of GaP and InP, calculated by using this large-basis method with a 65-element plane wave basis set and the atomic form factors of Table 15.7. The validity of this approach, which turns out to be a small generalisation for the bulk case, is substantiated by the close agreement obtained with the original calculations of Cohen and Bergstresser [4].

15.10 Spin–orbit coupling within the large-basis approach

In constructing spin-dependent solutions the basis set must contain spin-dependent (s or s') terms, hence the matrix elements given in equation (15.18), i.e.

$$H_{\mathbf{G}',\mathbf{G}} = \int u^*_{\mathbf{G}',\mathbf{k}} \mathcal{H} u_{\mathbf{G},\mathbf{k}} \, \mathrm{d}\tau \tag{15.79}$$

become:

$$H_{(\mathbf{G}',s'),(\mathbf{G},s)} = \int u^*_{\mathbf{G}',\mathbf{k},s'} \mathcal{H} u_{\mathbf{G},\mathbf{k},s} \, \mathrm{d}\tau \tag{15.80}$$

The additional potential V_{SO}, representing the spin–orbit interaction in the total Hamiltonian $\mathcal{H}$, leads to an additional term within each matrix element [7, 23, 24]. Under the auspices of the large-basis approach, this term is again a sum over the contributions from all the N_a atoms within the basis, and, by analogy with the crystal potential of equation (15.76), is given by:

$$V_{\mathrm{SO}} = \frac{1}{N_a} \sum_t \mathrm{e}^{-i\mathbf{q}_\bullet \mathbf{t}} \left[-i\lambda^{\mathbf{r}_a} \left(\mathbf{G}' + \mathbf{k} \right) \times (\mathbf{G} + \mathbf{k})_\bullet \sigma_{s',s} \right] \tag{15.81}$$

where $\lambda^{\mathbf{r}_a}$ is a parameter quantifying the magnitude of the interaction and $\mathbf{q} = \mathbf{G}' - \mathbf{G}$. The superscript $\mathbf{r}_a$ indicates that λ is dependent upon the atomic species at all the basis sites within the unit cell. The entity $\sigma_{s',s} = \langle s'|\sigma|s\rangle$, where:

$$\sigma = \sigma_1 \hat{\mathbf{i}} + \sigma_2 \hat{\mathbf{j}} + \sigma_3 \hat{\mathbf{k}} \tag{15.82}$$

and σ_1, σ_2 and σ_3 are the Pauli spin matrices:

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{15.83}$$

Thus:

$$\sigma = \begin{pmatrix} \hat{\mathbf{k}} & \hat{\mathbf{i}} - i\hat{\mathbf{j}} \\ \hat{\mathbf{i}} + i\hat{\mathbf{j}} & -\hat{\mathbf{k}} \end{pmatrix} \tag{15.84}$$

As partial analytical verification for this representation presented here, consider the case of a zinc blende basis consisting of two atoms, one cation and one anion. In this example, N_a would be 2 and $\mathbf{t} = \pm\mathbf{T}$, where, again, $\mathbf{T} = \frac{A_0}{8}(\hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}})$, i.e.

$$V_{\mathrm{SO}} = \frac{1}{2} \left[(-i\lambda^{\mathrm{cat}}) \mathrm{e}^{i\mathbf{q}_\bullet \mathbf{T}} + (-i\lambda^{\mathrm{an}}) \mathrm{e}^{-i\mathbf{q}_\bullet \mathbf{T}} \right] \left(\mathbf{G}' + \mathbf{k} \right) \times (\mathbf{G} + \mathbf{k})_\bullet \sigma_{s',s} \tag{15.85}$$

which gives:

$$V_{\mathrm{SO}} = \frac{1}{2}\left[-i\lambda^{\mathrm{cat}}\left(\cos\mathbf{q}_{\bullet}\mathbf{T} + i\sin\mathbf{q}_{\bullet}\mathbf{T}\right) - i\lambda^{\mathrm{an}}\left(\cos\mathbf{q}_{\bullet}\mathbf{T} - i\sin\mathbf{q}_{\bullet}\mathbf{T}\right)\right]$$

$$\times\left(\mathbf{G}' + \mathbf{k}\right) \times \left(\mathbf{G} + \mathbf{k}\right)_{\bullet}\sigma_{s',s} \tag{15.86}$$

Gathering terms in cos and sin:

$$V_{\mathrm{SO}} = \frac{1}{2}\left[-i(\lambda^{\mathrm{cat}} + \lambda^{\mathrm{an}})\cos\mathbf{q}_{\bullet}\mathbf{T} - i(\lambda^{\mathrm{cat}} - \lambda^{\mathrm{an}})i\sin\mathbf{q}_{\bullet}\mathbf{T}\right]\left(\mathbf{G}' + \mathbf{k}\right) \times \left(\mathbf{G} + \mathbf{k}\right)_{\bullet}\sigma_{s',s} \tag{15.87}$$

which gives:

$$V_{\mathrm{SO}} = \left[-i\left(\frac{\lambda^{\mathrm{cat}} + \lambda^{\mathrm{an}}}{2}\right)\cos\mathbf{q}_{\bullet}\mathbf{T} + \left(\frac{\lambda^{\mathrm{cat}} - \lambda^{\mathrm{an}}}{2}\right)\sin\mathbf{q}_{\bullet}\mathbf{T}\right]\left(\mathbf{G}' + \mathbf{k}\right) \times \left(\mathbf{G} + \mathbf{k}\right)_{\bullet}\sigma_{s',s} \tag{15.88}$$

The half sum and half difference of the λ parameters are often labelled as symmetric λ^{S} and anti-symmetric λ^{A} contributions to the spin–orbit interaction, thus giving:

$$V_{\mathrm{SO}} = \left(-i\lambda^{\mathrm{S}}\cos\mathbf{q}_{\bullet}\mathbf{T} + \lambda^{\mathrm{A}}\sin\mathbf{q}_{\bullet}\mathbf{T}\right)\left(\mathbf{G}' + \mathbf{k}\right) \times \left(\mathbf{G} + \mathbf{k}\right)_{\bullet}\sigma_{s',s} \tag{15.89}$$

which is the same result as in Chelikowsky and Cohen [24].

15.11 Computational implementation

With the aim of a computational implementation of this extension to the pseudo-potential method,[1] first consider the new extended basis set. The addition of spin information means that the basis set of plane waves has to be doubled (one set for spin 'up' and one set for spin 'down'). The procedure is to split the new matrix vertically and horizontally into four blocks, each of which is the same size as the original matrix:

$$\left(\begin{array}{c|c} \begin{array}{c}\text{Block 1}\\ s = +\frac{1}{2}, s' = +\frac{1}{2}\end{array} & \begin{array}{c}\text{Block 2}\\ s = +\frac{1}{2}, s' = -\frac{1}{2}\end{array} \\ \hline \begin{array}{c}\text{Block 3}\\ s = -\frac{1}{2}, s' = +\frac{1}{2}\end{array} & \begin{array}{c}\text{Block 4}\\ s = -\frac{1}{2}, s' = -\frac{1}{2}\end{array} \end{array}\right) \tag{15.90}$$

Each block has exactly the same crystal potential components as the original matrix, but with the additional spin-dependent component given in equation (15.81). A Kronecker delta $\delta_{s',s}$ acts upon the original components of each matrix element just to ensure that the only spin-dependent part is the new term V_{SO}. Thus the complete matrix elements, including the spin–orbit interaction within the large-basis approach, are given by:

$$H_{(\mathbf{G}',s'),(\mathbf{G},s)} = H_{\mathbf{G}',\mathbf{G}}\delta_{s',s} + V_{\mathrm{SO}} \tag{15.91}$$

[1] The author would like to thank Fei Long for his assistance.

where, for the large-basis approach, $H_{\mathbf{G}',\mathbf{G}}$ is given by equation (15.77). In full, this would be:

$$H_{(\mathbf{G}',s'),(\mathbf{G},s)} = \frac{\hbar^2}{2m_0}|\mathbf{G}+\mathbf{k}|^2\delta_{\mathbf{G}',\mathbf{G}}\delta_{s',s} + \frac{2}{N_a}\sum_t \mathrm{e}^{-i\mathbf{q}_\bullet\mathbf{t}}V_{\mathrm{f}}^{\mathbf{r}_a}(q)\delta_{s',s} + \frac{1}{N_a}\sum_t \mathrm{e}^{-i\mathbf{q}_\bullet\mathbf{t}}\left[-i\lambda^{\mathbf{r}_a}\left(\mathbf{G}'+\mathbf{k}\right)\times(\mathbf{G}+\mathbf{k})_\bullet\sigma_{s',s}\right] \tag{15.92}$$

Thus computationally the quickest way to generate this new extended matrix is to calculate the original crystal potential terms for just the lower triangle of the first 'block' and then copy it into the lower triangle of the fourth block. The spin–orbit term can then be added, with the only other **k**-dependent part, the kinetic energy term, added last along the leading diagonal.

For computational reasons it may be worthwhile considering the form of the spin–orbit contribution to the matrix elements, as this can be simplified for more efficient evaluation. With this in mind, consider the forms of $\sigma_{s',s}$. The spin states are represented by the spinors $(1,0)$ and $(0,1)$, hence:

$$\sigma_{\frac{1}{2},\frac{1}{2}} = (1,0)\begin{pmatrix}\hat{\mathbf{k}} & \hat{\mathbf{i}}-i\hat{\mathbf{j}}\\ \hat{\mathbf{i}}+i\hat{\mathbf{j}} & -\hat{\mathbf{k}}\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix} = (1,0)\begin{pmatrix}\hat{\mathbf{k}}\\ \hat{\mathbf{i}}+i\hat{\mathbf{j}}\end{pmatrix} = \hat{\mathbf{k}} \tag{15.93}$$

$$\sigma_{\frac{1}{2},-\frac{1}{2}} = (1,0)\begin{pmatrix}\hat{\mathbf{k}} & \hat{\mathbf{i}}-i\hat{\mathbf{j}}\\ \hat{\mathbf{i}}+i\hat{\mathbf{j}} & -\hat{\mathbf{k}}\end{pmatrix}\begin{pmatrix}0\\1\end{pmatrix} = (1,0)\begin{pmatrix}\hat{\mathbf{i}}-i\hat{\mathbf{j}}\\ -\hat{\mathbf{k}}\end{pmatrix} = \hat{\mathbf{i}}-i\hat{\mathbf{j}} \tag{15.94}$$

$$\sigma_{-\frac{1}{2},\frac{1}{2}} = (0,1)\begin{pmatrix}\hat{\mathbf{k}} & \hat{\mathbf{i}}-i\hat{\mathbf{j}}\\ \hat{\mathbf{i}}+i\hat{\mathbf{j}} & -\hat{\mathbf{k}}\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix} = (0,1)\begin{pmatrix}\hat{\mathbf{k}}\\ \hat{\mathbf{i}}+i\hat{\mathbf{j}}\end{pmatrix} = \hat{\mathbf{i}}+i\hat{\mathbf{j}} \tag{15.95}$$

$$\sigma_{-\frac{1}{2},-\frac{1}{2}} = (0,1)\begin{pmatrix}\hat{\mathbf{k}} & \hat{\mathbf{i}}-i\hat{\mathbf{j}}\\ \hat{\mathbf{i}}+i\hat{\mathbf{j}} & -\hat{\mathbf{k}}\end{pmatrix}\begin{pmatrix}0\\1\end{pmatrix} = (0,1)\begin{pmatrix}\hat{\mathbf{i}}-i\hat{\mathbf{j}}\\ -\hat{\mathbf{k}}\end{pmatrix} = -\hat{\mathbf{k}} \tag{15.96}$$

The vector product $\left(\mathbf{G}'+\mathbf{k}\right)\times(\mathbf{G}+\mathbf{k})$ is clearly equal to some other vector $A_x\hat{\mathbf{i}}+A_y\hat{\mathbf{j}}+A_z\hat{\mathbf{k}}$, say. Then in block 1 of the Hamiltonian matrix:

$$\left(\mathbf{G}'+\mathbf{k}\right)\times(\mathbf{G}+\mathbf{k})_\bullet\sigma_{s',s} = A_z \tag{15.97}$$

In block 2:

$$\left(\mathbf{G}'+\mathbf{k}\right)\times(\mathbf{G}+\mathbf{k})_\bullet\sigma_{s',s} = A_x - iA_y \tag{15.98}$$

In block 3:

$$\left(\mathbf{G}'+\mathbf{k}\right)\times(\mathbf{G}+\mathbf{k})_\bullet\sigma_{s',s} = A_x + iA_y \tag{15.99}$$

In block 4:

$$\left(\mathbf{G}'+\mathbf{k}\right)\times(\mathbf{G}+\mathbf{k})_\bullet\sigma_{s',s} = -A_z \tag{15.100}$$

15.12 Deducing the parameters and application

The primary aim of the book is to demystify the theory of semiconductor heterostructures, and make such calculations more *accessible*. The current topic of spin–orbit coupling within

the empirical pseudo-potential approach is one example where this is paramount, as the literature can be quite confusing. The spin–orbit interaction V_{SO} given in equation (15.81) is an energy, and λ is usually quoted in rydbergs (a unit of energy), thus the vectors $\mathbf{G}$ and $\mathbf{k}$ must be dimensionless. The question is what dimensions? In this work the result of the vector and scalar product, i.e. $\left(\mathbf{G}' + \mathbf{k}\right) \times \left(\mathbf{G} + \mathbf{k}\right)_{\bullet} \sigma_{s',s}$, is made dimensionless by dividing by $(2\pi/A_0)^2$. As the structure factor $\mathrm{e}^{-i\mathbf{q}_{\bullet}\mathbf{t}}$ is dimensionless, the contribution V_{SO} to the matrix element takes the units of the parameter λ.

Not all authors use this convention, and this is reflected in the literature by the usual choice to quote the spin–orbit splitting Δ_0 between the doubly degenerate light- and heavy-hole bands and the split-off band. In some instances the symmetric λ^{S} and anti-symmetric λ^{A} parameters are quoted for binary compounds, and it is a simple task to obtain the spin–orbit parameters for the individual atoms as required here:

$$\lambda^{\mathrm{cat}} = \lambda^{\mathrm{S}} - \lambda^{\mathrm{A}}, \quad \lambda^{\mathrm{an}} = \lambda^{\mathrm{S}} + \lambda^{\mathrm{A}} \tag{15.101}$$

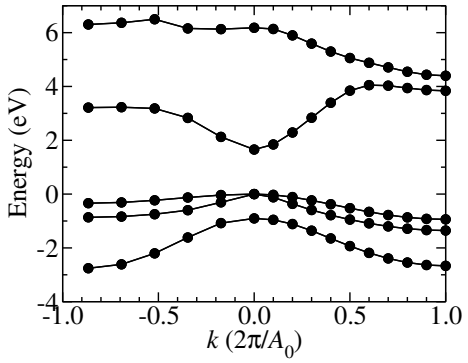


Figure 15.12: The band structure of bulk CdTe calculated with 169 plane waves and adjusted to set the top of the valence band to zero

Figure 15.12 shows the results of a calculation of the band structure of CdTe using the original form factors of Cohen and Bergstresser [4] with the spin–orbit parameters λ^{S} and λ^{A} quoted by Bloom and Bergstresser [8]. It was found that the λ values had to be multiplied by a factor of 0.343 in order to obtain a splitting Δ_0 of $905\,\mathrm{meV}$ in agreement with the values of 910 and $900\,\mathrm{meV}$ quoted by Bloom and Bergstresser. This is a relatively quick process because, to *first order*, the valence-band splitting is proportional to the λ parameters, so if the result for Δ_0 is half the experimental value, scaling the lambda parameters by a factor of 2 will 'get you close'. The splitting of the light and heavy hole at the X ($k = (0, 0, 2\pi/A_0)$) is $415\,\mathrm{meV}$ and at the L point ($k = (\pi/A_0, \pi/A_0, \pi/A_0)$) is $525\,\mathrm{meV}$, which are both in reasonable agreement with the values quoted by Chelikowsky and Cohen [24] of 380 and $530\,\mathrm{meV}$, respectively. These results serve to substantiate the validity of the spin–orbit extension. Note, though, that there is some discrepancy in the calculations here and disagreement between the results quoted [8, 24]. This serves as an illustration of the

uncertainty surrounding the knowledge of remote areas of the band structure (away from high symmetry points and band edges), but also it must be noted that the spin–orbit parameters should be chosen self-consistently with the atomic form factors—a tedious process that will not be attempted here.

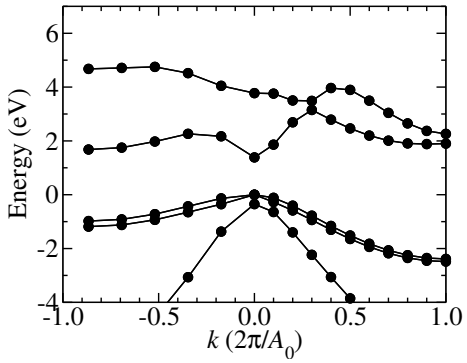


Figure 15.13: The band structure of bulk GaAs using the spin-free form factors of Mader and Zunger [25] together with the spin–orbit parameters described in the text

Taking μ' as some adjustable parameter (which can again absorb any differences in units employed), Chelikowsky and Cohen [24] give the spin–orbit splittings for a binary compound AB as:

$$\lambda^{\mathrm{A}} = \mu', \quad \lambda^{\mathrm{B}} = \alpha\mu' \tag{15.102}$$

where α is the ratio of the spin–orbit splittings of the free atoms. The latter can be looked up in a book [26]. Unfortunately, the book only gives data for atoms found in II–VI compounds, so it is difficult to know what α should be for any III–V material. Thus any calculation for a III–V material, such as GaAs, would be dependent upon two parameters, which means a lot of work may be required to limit the values they take and produce good agreement with experiment.

To 'cut a corner', consider just *interpolating* the spin–orbit splittings of the free Ga and As atoms from their neighbouring entries in the tabulated data of Herman and Skillman ([26], pp. 2–6), using their notation:

$$\mathrm{E(S\text{-}O)(GA)} = \frac{\mathrm{E(S\text{-}O)(Zn)+E(S\text{-}O)(Ge)}}{2} \tag{15.103}$$

$$\mathrm{E(S\text{-}O)(AS)} = \frac{\mathrm{E(S\text{-}O)(Ge)+E(S\text{-}O)(Se)}}{2} \tag{15.104}$$

i.e.

$$\mathrm{E(S\text{-}O)(GA)} = \frac{-0.00596 - 0.00515}{2} = -0.00556 \tag{15.105}$$

$$\mathrm{E(S\text{-}O)(AS)} = \frac{-0.00515 - 0.01026}{2} = -0.00771 \tag{15.106}$$

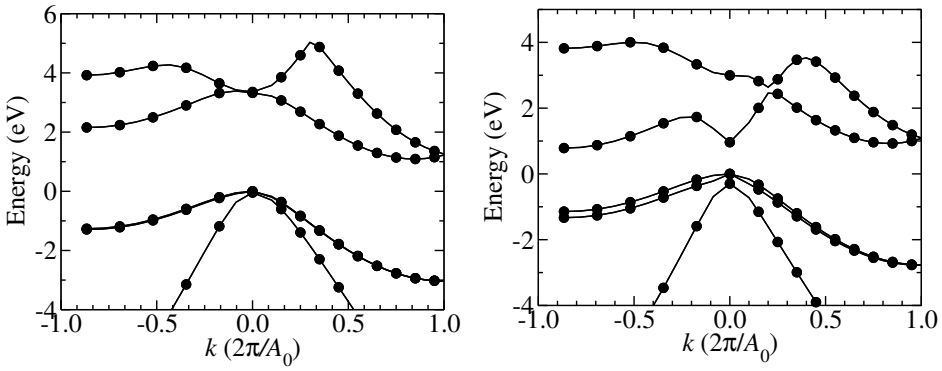


Figure 15.14: The band structures of bulk Si (a) and Ge (b), including spin–orbit coupling along the usual $\mathrm{L} \rightarrow \Gamma \rightarrow \mathrm{X}$ line

giving:

$$\alpha = \frac{\text{E(S-O)(AS)}}{\text{E(S-O)(GA)}} = 1.38 \tag{15.107}$$

Using this, the value of μ' was adjusted (and found to be 0.000402) to give a valence-band splitting of $350\,\mathrm{meV}$. This in itself is not enough to substantiate the calculation; however, at the same time the light-hole–heavy-hole splitting at X is $97\,\mathrm{meV}$, which is incredibly close to the $100\,\mathrm{meV}$ quoted by Chelikowsky and Cohen (see Fig. 15.13).

This shows the robustness of the method—while this GaAs calculation has been forced through and is not necessarily scientifically rigorous, this calculation does demonstrate that the technique can be used to obtain the valence-band structure of III–V materials to some degree of acceptability. Should further accuracy be required, more time could justifiably be invested in determining a better set of parameters.

It is a much simpler task to deduce the spin–orbit parameters of the elemental group IV semiconductors because, with only one atom type, there is only one parameter. Using the methods outlined above, λ was found to be 0.000106 and 0.00058 rydbergs, giving valence-band splittings of 44 and $296\,\mathrm{meV}$ for Si and Ge, respectively. This gave rise to the band structures plotted in Fig. 15.14. The calculations are based on the spin-free atomic form factors for Si and Ge heterostructures given by Friedel *et al*. [27]. They are in excellent agreement with the data presented by Chelikowsky and Cohen and, in particular for Ge, the light-hole–heavy-hole splitting at the L-point is $186\,\mathrm{meV}$, in comparison with the quoted value of $200\,\mathrm{meV}$.

15.13 Isoelectronic impurities in bulk

A generalisation of the standard two-atom basis within the empirical pseudo-potential method was demonstrated in Section 15.9, although hitherto the calculations presented have merely seemed more complex than previously and the investment seems to have produced

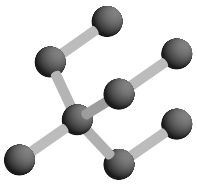


Figure 15.15: The non-primitive eight-atom basis of a group IV face-centred cubic crystal (diamond structure)

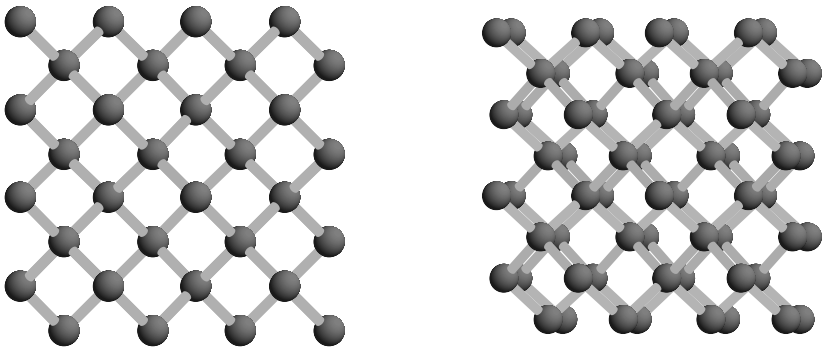


Figure 15.16: A 64-atom basis for the diamond structure, shown looking along one of the $\langle 100 \rangle$ directions (left) and slightly off-axis (right)

no visible benefit. In fact, the generalisation was motivated by the desire to study more complex distributions of atoms within a crystal, where primarily this implies quantum well heterostructures, wires and dots (considered in the next chapter). However, in addition, some bulk systems of interest still remain and the techniques are now at hand for their solutions.

The first case in question is that of a single impurity atom within a host semiconductor. This topic is itself extensive, and has already been given a substantial chapter in this text within the envelope function and effective mass approximations. The interest now lies in the ability to explore the microscopic electronic properties from a more fundamental point of view. Indeed, such a detailed study has already been undertaken by Jaros [2]—here, a very brief demonstration will be given as to how the large-basis calculation *may* be applied to such instances.

The first point to be addressed concerns the nature of the atomic basis. Clearly, a host crystal needs to be specified and a single atom, representing the impurity, needs to be placed substitutionally on one of the lattice sites. It is perhaps convenient in this scenario to consider a cube of Si, and in this case the impurity atom chosen will be Ge, which would, of course, be *isoelectronic*.

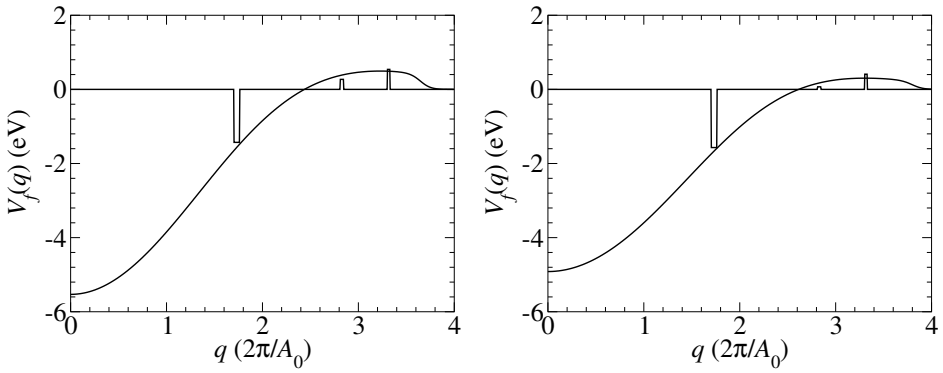


Figure 15.17: The discrete (δ-function-like) pseudo-potential form factors of the face-centred cubic lattice [4] compared with the continuous form necessary for systems of reduced dimensionality, for Si (left) and Ge (right) [27]

Ashcroft and Mermin ([3], p. 75), give the non-primitive face-centred cubic four-point basis vectors as:

$$\mathbf{T}_0 = \mathbf{0}, \quad \mathbf{T}_1 = \frac{A_0}{2}(\hat{\mathbf{i}} + \hat{\mathbf{j}}), \quad \mathbf{T}_2 = \frac{A_0}{2}(\hat{\mathbf{j}} + \hat{\mathbf{k}}), \quad \mathbf{T}_3 = \frac{A_0}{2}(\hat{\mathbf{k}} + \hat{\mathbf{i}}) \tag{15.108}$$

These are chosen to represent the cubic symmetry of the face-centred cubic lattice more clearly. When applied to the diamond or zinc blende structures, such a set of vectors would then imply an eight-atom basis, with atoms at $\pm\mathbf{T}=\pm\frac{A_0}{8}(\hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}})$ for each of the above, i.e.

$$\mathbf{r}_a = \mathbf{T}_i \pm \frac{A_0}{8}(\hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}}), \quad i = 0, 1, 2, 3 \tag{15.109}$$

as displayed in Fig. 15.15. The Bravais lattice vectors then simply follow as:

$$\mathbf{a}_1 = A_0\hat{\mathbf{i}}, \quad \mathbf{a}_2 = A_0\hat{\mathbf{j}}, \quad \mathbf{a}_3 = A_0\hat{\mathbf{k}} \tag{15.110}$$

Such an alternative way of constructing the crystal is useful for defining large cells within a host semiconductor. In this case, where a description of a single isolated impurity is the aim, it is necessary to generalise this to produce a larger cube. For example, if n_x, n_y and n_z of these cubes were stacked together along the x-, y-, and z-axes, respectively, then the atomic positions within this larger basis would be given by:

$$\mathbf{r}_a = m_x A_0\hat{\mathbf{i}} + m_y A_0\hat{\mathbf{j}} + m_z A_0\hat{\mathbf{k}} + \mathbf{T}_i \pm \frac{A_0}{8}(\hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}}) \tag{15.111}$$

where again the index $i = 0$, 1, 2 and 3 and the additional indices $m_{x,y,z} = 0, 1, 2, \ldots, n_{x,y,z}$. Figure 15.16 illustrates a large basis produced by stacking eight face-centred cubic cells together, i.e. $n_x = n_y = n_z = 2$, thus giving 64 atoms in total.

For this large ($n_x \times n_y \times n_z$) cuboid, the primitive reciprocal lattice vectors would simply generalise to:

$$\mathbf{b}_1 = \frac{1}{n_x}\frac{2\pi}{A_0}\hat{\mathbf{i}}, \quad \mathbf{b}_2 = \frac{1}{n_y}\frac{2\pi}{A_0}\hat{\mathbf{j}}, \quad \mathbf{b}_3 = \frac{1}{n_z}\frac{2\pi}{A_0}\hat{\mathbf{k}} \tag{15.112}$$

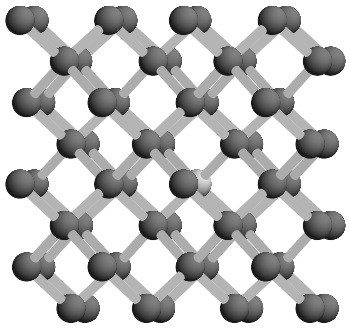


Figure 15.18: A 64-atom basis for the diamond structure, with a single substitutional Ge impurity (lighter colour, near centre)

Hence, the basis **G** set for expansion is given by all linear combinations of the above, i.e.

$$\mathbf{G} = \beta_1 \mathbf{b}_1 + \beta_2 \mathbf{b}_2 + \beta_3 \mathbf{b}_3, \quad \beta_i \in \mathcal{Z} \tag{15.113}$$

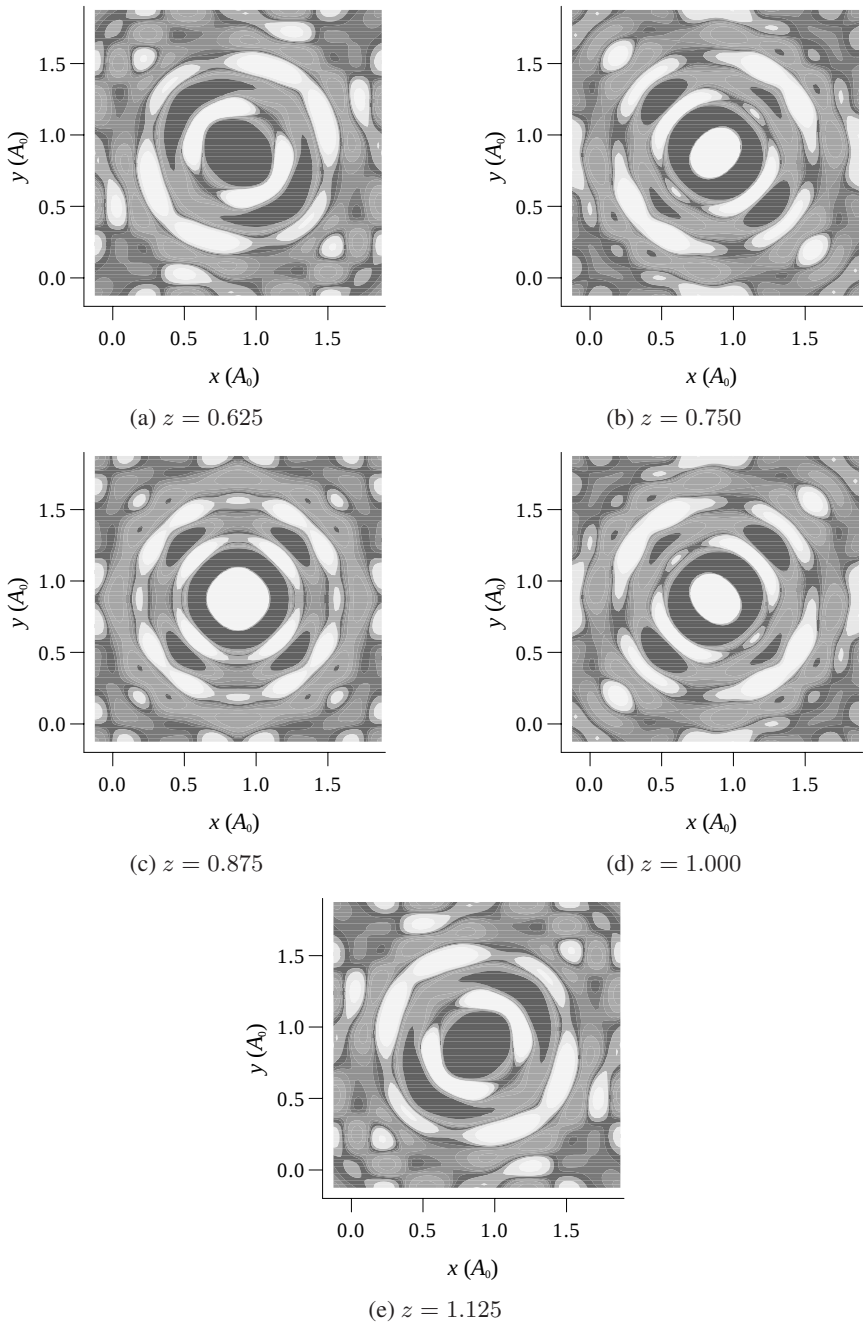
Note that this set of reciprocal lattice vectors implies that the magnitude of the difference vector, $\mathbf{q} = \mathbf{G}' - \mathbf{G}$, can take many more values than the discrete set of the face-centred cubic structure (given in Table 15.1). For this reason (and many others), more recently deduced empirical pseudo-potential form factors are often represented by a continuous curve rather than by a discrete set of points. Figure 15.17 illustrates this for the Si and Ge form factors of Friedel *et al*. [27], and compares them with the original discrete form factors of Cohen and Bergstresser [4], employed earlier in this chapter.

It can be seen that the agreement between the two representations of the pseudo-potential is close and both have been truncated at $q = 4$, although at the higher end of the domain there is a slight difference. The continuous potentials of Friedel *et al*. [27] have been deduced by comparison with a variety of experimental results and have been shown to reproduce a large amount of experimental data, including bulk band structure, deformation potentials, and heterostructure band offsets. Truncating the reciprocal lattice vectors of equation (15.112) at the same point as the potential ($q = 4$) gives a basis set of 2109 plane waves. Hence, the Hamiltonian matrix H has 2109^2 elements, each of which is a complex number. Using double-precision arithmetic (8 bytes per number), this implies the Hamiltonian matrix will occupy $2109^2 \times 2 \times 8$ bytes ≈ 71 MB, which is well within the reach of a standard desktop computer. However, it can be seen that the memory demands for this type of calculation can grow very quickly and this is one of the detrimental points of the large-basis method. While it is very simple, convenient and flexible, simply employing direct diagonalisation of the matrix as the computational solution is quite inefficient. However, a calculation *of this size* is manageable and all is now in place to proceed with it.

In this analysis one of the 64 Si atoms in the basis is substituted by Ge, as shown in Fig. 15.18. The choice of which atom is not important, as the periodic nature of the pseudo-potential calculations ensures that they are all equivalent. Just as the two-atom basis of the simple elemental Si calculation produces four valence-band states, this more complex 64-atom basis gives $64 \times 2 = 128$ valence-band states. Figure 15.19 shows the change in the

total valence-band charge density on substitution of the single Ge isoelectronic impurity at an electron wave vector $\mathbf{k} = 0$; note that the lighter the colour, the larger the increase.

In this particular example, the Si atom with coordinates $(0.875, 0.875, 0.875)A_0$ was substituted, and hence the third plane, with $z = 0.875A_0$, intersects the centre of the Ge atom. Focusing on this plot, the charge density on the Ge atomic site is higher than that on the surrounding crystal and is fairly well localised, which (just about) justifies the initial choice of a 64-atom basis. The remaining four figures display the change in the charge density above and below the plane of this isolated Ge atom, where the symmetry is apparent equidistant from the centre of the impurity, and it can be seen that a quarter of a lattice constant away (1.4 Å), as shown on the first and fifth planes, the charge density has returned to its background (Si) value.



(a) $z = 0.625$ (b) $z = 0.750$

(c) $z = 0.875$ (d) $z = 1.000$

(e) $z = 1.125$

Figure 15.19: The change in the total valence-band charge density (ignoring spin–orbit coupling) around a single Ge atom within a Si host across the x–y plane of four face-centred cubic unit cells, for a variety of z values (given in A_0); note that the lighter the colour, the higher the increase in the charge density

15.14 The electronic structure around point defects

Such an approach can also be used to model point defects. For example, instead of substituting a Si atom for an isoelectronic Ge atom, the atom can be removed completely to represent a vacancy. Now this is a simplified picture, in that for this example it is assumed that all of the other atoms remain on their diamond lattice sites. This, of course, will not be the case, as there will be some localised relaxation of the crystal around the vacancy, i.e. some movement of the Si atoms.

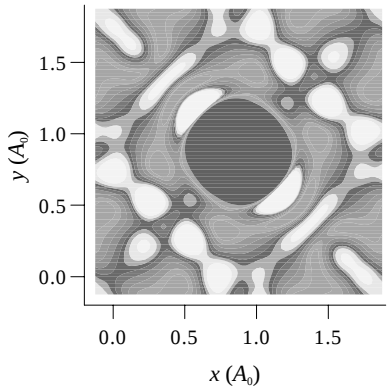


Figure 15.20: The change in the total valence-band charge density (ignoring spin–orbit coupling) around a vacancy within a Si host across the x–y plane of four face-centred cubic unit cells; note that the lighter the colour, the higher the increase in the charge density

Again employing the 64-atom cubic basis as above, Fig. 15.20 displays the change in the electronic structure around a single vacancy when compared to the unperturbed Si host crystal. The plot centres the vacancy site and represents a cross-section of four (2×2) face-centred cubic unit cells. The absence of charge is clearly illustrated by the darker shading, and interestingly suggests a 'halo' of increased charge density around the vacancy. Such a movement of charge will aid shielding of the 'hole' from the remainder of the crystal.

Exercises

(1). Starting from a knowledge of the Bravais lattice vectors, derive the reciprocal lattice vectors for:

(a) a simple cubic crystal;

(b) a body-centred cubic Bravais lattice.

Comment on the results.

(2). The choice of the origin at the centre of the axis between the two Si (or Ge) atoms in the elemental bandstructure calculation leads to a particularly simple form for the structure factor S:

$$S = N \sum_{\mathbf{t}} \mathrm{e}^{i(\mathbf{G}-\mathbf{G}')\bullet\mathbf{t}} = 2N \cos{(\mathbf{G}-\mathbf{G}')}_\bullet \mathbf{T}$$

Deduce the new form of the structure factor S and comment on the effect on the reciprocal lattice vectors **G** if:

(a) the origin is shifted on to one of the atoms within the two-atom basis;

(b) the axes are rotated such that the two-atom basis lies along the x-axis.

(3). Just three form factors (see Table 15.2) characterise the Si band structure. Using the `qwwad_pp_large_basis` program, perform a sensitivity analysis by varying each of the form factors in turn to see the effect on the band structure. Focus, in particular, on the energy minimum in the valley near the X-point in the Brillouin zone.

(4). Use the `qwwad_pp_large_basis` program to calculate the electron effective mass in silicon in the valley near the X-point of the Brillouin zone.

(5). Use the `qwwad_pp_large_basis` program to calculate to a high degree of accuracy the band structure around the Γ-point in GaAs, and hence deduce the deviation from a parabolic dispersion (E–**k**) curve.

(6). The pseudo-potential method can be used to describe heterojunctions of different materials. This is achieved by adjusting the form factors for each material such that the energy difference between the tops of the valence bands agrees with that measured by experiment. First, considering GaAs in isolation, adjust the $V_{\mathrm{f}}(q=0)$ form factor in order to give the valence band at Γ equal to zero when calculated with the `qwwad_pp_large_basis` program. Find form factors for AlAs and repeat to give the top of the valence band at $-0.412\,\mathrm{meV}$. Plot both band structures on the same figure.

(7). Calculate the band structure of a range of alloy concentrations x of $\mathrm{Ga}_{1-x}\mathrm{Al}_x\mathrm{As}$ and hence show how the band gap of GaAs goes from being direct to indirect as x increases beyond 0.4. Hint: use the `qwwad_pp_large_basis` program with the atom types 'ALGAAS' and 'ASALGA'; note that the Al concentration x can be included into these atom types by including two additional numbers (e.g. for $x = 0.2$ the atom types would be 'ALGAAS20' and 'ASALGA20').

References

[1] W. A. Harrison, *Pseudopotentials in the Theory of Metals*, W. A. Benjamin, New York, 1966.

[2] M. Jaros, *Deep Levels in Semiconductors*, Hilger, Bristol, 1982.

[3] N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Saunders College Publishing, Philadelphia, 1976.

[4] M. L. Cohen and T. K. Bergstresser, 'Band structures and pseudopotential form factors for fourteen semiconductors of the diamond and zinc-blende structures', *Phys. Rev.*, **141**:789, 1966.

[5] G. Bastard, *Wave Mechanics Applied to Semiconductor Heterostructures*, Monographies de physique. Halsted Press, New York, 1988.

[6] R. M. Eisberg, *Fundamentals of Modern Physics*, John Wiley & Sons, Inc., New York, 1961.

[7] G. Weisz, 'Band structure and Fermi surface of white tin', *Phys. Rev.*, **149**:504, 1966.

[8] S. Bloom and T. K. Bergstresser, 'Band structure of α-Sn, InSb and CdTe including spin-orbit effects', *Solid State Commun.*, **6**(7):465–467, 1968.

[9] S. Adachi, *Physical Properties of III–V Semiconductor Compounds: InP, InAs, GaAs, GaP, InGaAs, and InGaAsP*, John Wiley & Sons, Inc., New York, 1992.

[10] L. A. Coldren and S. W. Corzine, *Diode Laser and Photonic Integrated Circuits*, John Wiley & Sons, Inc., New York, 1995.

[11] S. Nakamura, M. Senoh, N. Iwasa, S.-i. Nagahama, T. Yamada, and T. Mukai, 'Superbright green InGaN single-quantum-well-structure light-emitting diodes', *Jpn. J. Appl. Phys.*, **34**(10B):L1332, 1995.

[12] S. Nakamura, M. Senoh, S. Nagahama, N. Iwasa, T. Yamada, T. Matsushita, Y. Sugimoto, and H. Kiyoku, 'Continuous-wave operation of InGaN multi-quantum-well-structure laser diodes at 233 K', *Appl. Phys. Lett.*, **69**:3034, 1996.

[13] S. Nakamura, 'Characteristics of room temperature-CW operated InGaN multi-quantum-well-structure laser diodes', *MRS Internet J. Nitride Semicond. Res.*, **2**, 1997.

[14] J. S. Blakemore, *Solid State Physics*, Cambridge University Press, Cambridge, Second edition, 1985.

[15] Y. C. Yeo, T. C. Chong, and M. F. Li, 'Electronic band structures and effective mass parameters of wurtzite GaN and InN', *J. Appl. Phys.*, **83**:1429, 1998.

[16] S. Adachi, *GaAs and Related Materials: Bulk Semiconducting and Superlattice Properties*, World Scientific, Singapore, 1994.

[17] P. Harrison, F. Long, and W. E. Hagston, 'Empirical pseudo-potential calculation of the in-plane effective masses of electron and holes of two-dimensional excitons in CdTe quantum wells', *Superlatt. Microstruct.*, **19**:123–130, 1996.

[18] S. C. Jain, *Germanium-Silicon Strained Layers and Heterostructures*, Advances in Electronics and Electron Physics. Academic Press, Boston, 1994.

[19] A.-B. Chen and A. Sher, *Semiconductor Alloys: Physics and Materials Engineering*, Plenum, New York, 1995.

[20] G. P. Strivastava, J. L. Martins, and A. Zunger, 'Atomic structure and ordering in semiconductor alloys', *Phys. Rev. B*, **31**:2561, 1985.

[21] S.-H. Wei and A. Zunger, 'Band gap narrowing in ordered and disordered semiconductor alloys', *Appl. Phys. Lett.*, **56**:662, 1990.

[22] S.-H. Wei and A. Zunger, 'Optical properties of zinc blende alloys: effects of epitaxial strain and atomic ordering', *Phys. Rev. B*, **49**:14337, 1994.

[23] J. P. Walter and M. L. Cohen, 'Calculated and measured reflectivity of ZnTe and ZnSe', *Phys. Rev. B*, **1**:2661, 1970.

[24] J. R. Chelikowsky and M. L. Cohen, 'Nonlocal pseudopotential calculations for the electronic structure of eleven zinc-blende semiconductors', *Phys. Rev. B*, **14**:556, 1976.

[25] K. A. Mäder and A. Zunger, 'Empirical atomic pseudopotentials for AlAs/GaAs superlattices, alloys, and nanostructures', *Phs. Rev. B*, **50**:17393, 1994.

[26] F. Herman and S. Skillman, *Atomic Structure Calculations*, Prentice Hall, Englewood Cliffs, NJ, 1963.

[27] P. Friedel, M. S. Hybertsen, and M. Schlüter, 'Local empirical pseudopotential approach to the optical properties of Si/Ge superlattices', *Phys. Rev. B*, **39**(11):7974–7977, 1989.

16

Pseudo-potential calculations of nanostructures

16.1 The superlattice unit cell

The large-basis approach to pseudo-potential calculation can be applied to semiconductor heterostructures, where, in particular, the short-period nature of the repeating unit cell in a superlattice is an ideal example. In order to proceed with such a calculation, it is necessary only to deduce the Bravais lattice vectors describing the crystal symmetry of the superlattice and from these to derive the reciprocal lattice vectors.

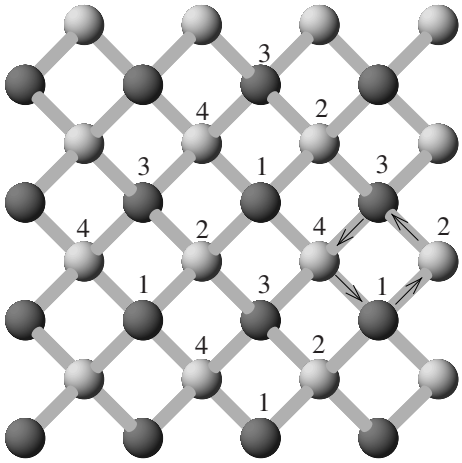


Figure 16.1: Looking along one of the $\langle 100 \rangle$ axes of a zinc blende crystal lattice

Figure 16.1 displays the now familiar zinc blende crystal lattice along one of the $\langle 100 \rangle$ directions. The crystal can be thought of as being composed of a series of 'one-dimensional' spirals that project into the plane of the paper. Such a spiral begins on an atom (in this case a cation) labelled '1' and then proceeds anticlockwise through atoms '2', '3' and '4', with the fifth atom in the spiral being exactly one lattice constant behind the first. This is illustrated

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

more clearly in Fig. 16.2. Choosing the z-axis along the length of the spiral, the atomic positions, in units of A_0, follow as:

$$\mathbf{r}_1 = \left(-\frac{1}{8}, -\frac{1}{8}, -\frac{1}{8}\right) \tag{16.1}$$

$$\mathbf{r}_2 = \left(\frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right) \tag{16.2}$$

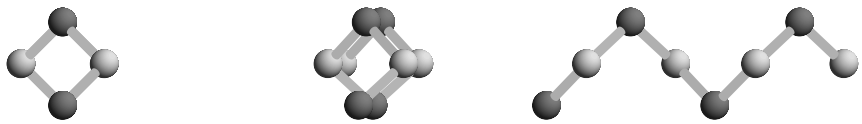
$$\mathbf{r}_3 = \left(-\frac{1}{8}, \frac{3}{8}, \frac{3}{8}\right) \tag{16.3}$$

$$\mathbf{r}_4 = \left(-\frac{3}{8}, \frac{1}{8}, \frac{5}{8}\right) \tag{16.4}$$

The spiral then continues with a period (in this bulk case) of $A_0\hat{\mathbf{k}}$, and thus:

$$\mathbf{r}_5 = \mathbf{r}_1 + (0, 0, 1) \tag{16.5}$$

$$\mathbf{r}_6 = \mathbf{r}_2 + (0, 0, 1) \tag{16.6}$$



(a) (b) (c)

Figure 16.2: A single spiral of zinc blende: (a) looking along one of the $\langle 100 \rangle$ directions; (b) slightly off-axis; (c) side view

This is easily generalised to a superlattice by simply defining the atomic positions with the appropriate choice of species. Figure 16.3 illustrates this for a diamond (just to keep the illustration simple) heterostructure. Again, the atomic positions within the unit cell are repetitions of the four primitive vectors $\mathbf{r}_1, \ldots, \mathbf{r}_4$ given in equations (16.1)–(16.4). Thus for a superlattice of period $n_z A_0$, the position vectors of the atoms within this larger basis are given by:

$$\mathbf{t} = \mathbf{r}_i + i_{n_z} A_0 \hat{\mathbf{k}}, \quad i = 1, 2, 3, 4;\ i_{n_z} = 0, 1, 2, \ldots, n_z - 1 \tag{16.7}$$

The Bravais lattice vector describing the periodicity of the superlattice along the axis of symmetry (usually defined as the z-axis) is therefore clearly just $n_z A_0 \hat{\mathbf{k}}$. Referring back to Fig. 16.1, the Bravais lattice vectors describing the in-plane (x–y) periodicity are those linking the atoms labelled '1', i.e. in units of A_0, $(\frac{1}{2}, \frac{1}{2}, 0)$ and $(\frac{1}{2}, -\frac{1}{2}, 0)$. In summary, the

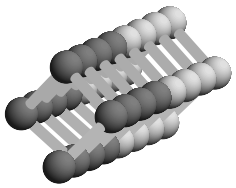


Figure 16.3: Off-axis view along the single spiral unit cell of a diamond heterostructure

Bravais lattice vectors for a superlattice are:

$$\mathbf{a}_1 = A_0 \left(\frac{1}{2}\hat{\mathbf{i}} + \frac{1}{2}\hat{\mathbf{j}} \right) \tag{16.8}$$

$$\mathbf{a}_2 = A_0 \left(\frac{1}{2}\hat{\mathbf{i}} - \frac{1}{2}\hat{\mathbf{j}} \right) \tag{16.9}$$

$$\mathbf{a}_3 = n_z A_0 \hat{\mathbf{k}} \tag{16.10}$$

The primitive reciprocal lattice vectors then follow from the usual vector product relationships (see Chapter 1), as:

$$\mathbf{b}_1 = \frac{2\pi}{A_0} \left(\hat{\mathbf{i}} + \hat{\mathbf{j}} \right) \tag{16.11}$$

$$\mathbf{b}_2 = \frac{2\pi}{A_0} \left(\hat{\mathbf{i}} - \hat{\mathbf{j}} \right) \tag{16.12}$$

$$\mathbf{b}_3 = \frac{2\pi}{A_0} \frac{1}{n_z} \hat{\mathbf{k}} \tag{16.13}$$

The complete set of reciprocal lattice vectors required for the pseudo-potential calculation is created by taking linear combinations of the above primitive vectors. Although this could be worked out analytically, it is quicker to take the short-cut of generating this set numerically with a simple computer program. In this manner, Fig. 16.4 plots the number of plane waves in the expansion set (when limiting the maximum magnitude to $4 \times 2\pi/A_0$) as a function of the superlattice period.

A linear fit to the data shown in Fig. 16.4 gives the number of plane waves as:

$$N \approx 128 n_z \tag{16.14}$$

which, as can be seen, can become quite large for typical periods. However, for a short-period superlattice, $n_z = 10A_0$ corresponds to 56.5 Å in the GaAs/$Ga_{1-x}Al_xAs$ system. This simplistic approach generates 1281 plane waves, thus giving a Hamiltonian matrix $\mathsf{H}_{\mathbf{G}',\mathbf{G}}$ of order 1281. With each element being a complex ($2\times$) double (8 bytes), this matrix requires $2 \times 8 \times 1281^2 \sim 26$ MB of storage. A quantum cascade laser with a period of 250 Å would have $n_z = 50A_0$, say, giving a Hamiltonian matrix of order 6405 and a memory requirement of ~ 650 MB.

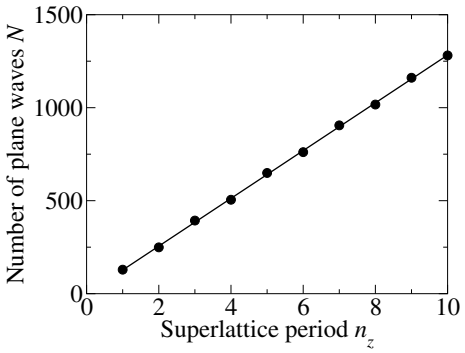


Figure 16.4: The number of plane waves required in the expansion set for the large-basis approach to superlattices

16.2 Application of large-basis method to superlattices

While not perhaps being the most computationally 'slick' method, the large-basis approach to the empirical pseudo-potential calculation of an extended crystallographic unit cell is straightforward and offers an excellent introduction to such calculations for researchers new to the field. In the previous section, the application of this technique was introduced for a particular classification of semiconductor heterostructures, namely superlattices with a relatively short period. In these first example calculations, this period will be considered to be constructed from a single layer of GaAs coupled to a single layer of AlAs, as illustrated schematically in Fig. 16.5.

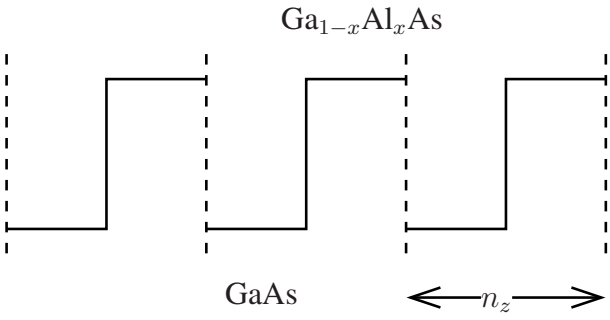


Figure 16.5: Schematic illustration of the repeating unit cell, of length $n_z A_0$, within a superlattice

As mentioned often before, a knowledge of the pseudo-potentials for superlattices is required at many more points than the small discrete set akin to bulk materials. With this aim, the pseudo-potentials of Mäder and Zunger [1], which have been shown to reproduce

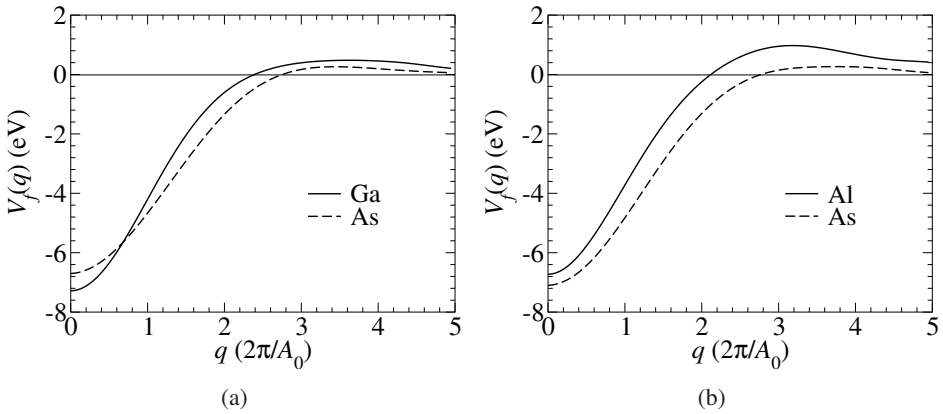


Figure 16.6: The continuous atomic pseudo-potentials of (a) GaAs and (b) AlAs of Mäder and Zunger [1]

many experimentally measured parameters as well as the results of *ab initio* calculations, will be employed. The required continuous nature of these potentials is illustrated in Fig. 16.6 for both atomic species in GaAs and AlAs.

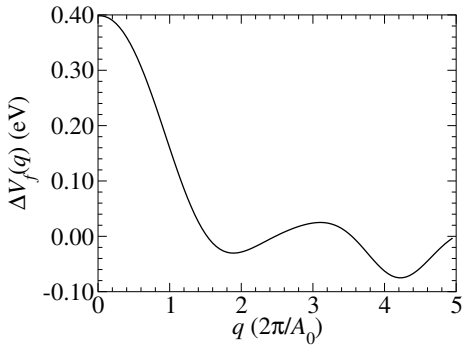


Figure 16.7: The difference in the atomic potentials of As in GaAs and AlAs

Again, notice how the atomic potential of As in GaAs is different from that in AlAs. This difference, $\Delta V_{\rm f}(q) = V_{\rm f}^{\rm As\ in\ GaAs}(q) - V_{\rm f}^{\rm As\ in\ AlAs}(q)$, is plotted in Fig. 16.7, just to re-emphasise the point made in Section 15.8, namely that, unfortunately, universal atomic potentials do not actually exist.

Figure 16.8 shows the band structures of bulk GaAs and AlAs, *ignoring spin–orbit coupling*, as produced from the continuous potentials. Both band structures have been translated upwards in energy such that the top of the valence band of GaAs is at zero. At the same time, the top of the valence band for AlAs is at $-0.51\,{\rm eV}$, which gives the

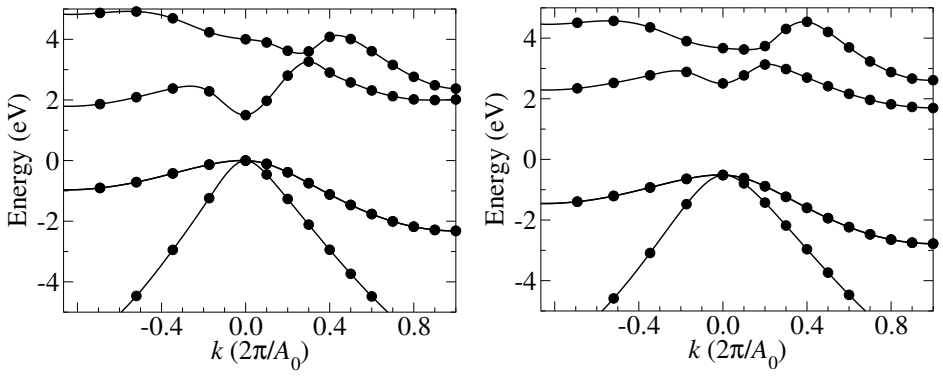


Figure 16.8: The bulk band structures of (a) GaAs and (b) AlAs, illustrating the valence-band offset built into the potentials of Mäder and Zunger [1]

offset in the valence band as 33% of the total, in line with experimentally determined values (see, for example, [2], p. 179), and recent calculations (see, for example, [3]). This illustrates an important point, namely that the correct choice of potentials not only can just be used to reproduce experimental data for bulk material, but also can retain information on the alignment (or band offset) of two dissimilar materials.

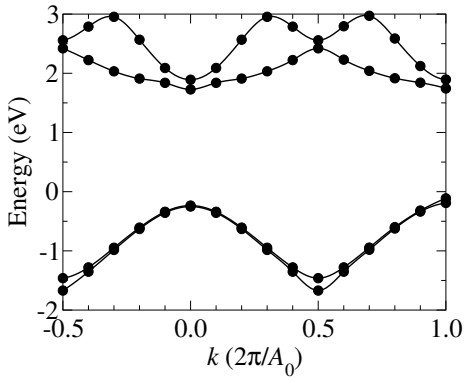


Figure 16.9: The band structure of a 1 ML GaAs/1 ML AlAs superlattice around the fundamental gap

Therefore, all is in place for the calculation to proceed: the superlattice unit cell has been defined, the reciprocal lattice vectors have been calculated, and the pseudo-potentials have been shown to reproduce the bulk band structure *and* the band alignment between the materials of interest. Figure 16.9 displays the results of calculations of the energy-level

structure around the fundamental gap of a 1 ML (monolayer) GaAs/1 ML AlAs superlattice (sometimes written as $(GaAs)_1(AlAs)_1$), i.e. $n_z = 1$, for a range of electron wave vectors k along the z-axis (parallel to the growth direction). The cyclic nature of the dispersion curves is evident with a period of $2\pi/A_0$, which represents the primitive reciprocal lattice vector along that direction, as given in equation (16.13). Thus, the edges of the superlattice Brillouin zone are at $k = \pm\pi/(n_z A_0)$, which is again the commonly found result.

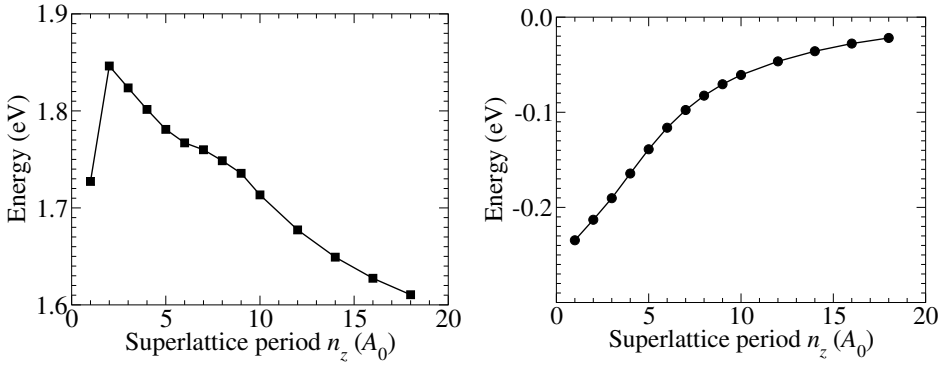


Figure 16.10: The energies of the (a) lowest conduction-band state and (b) highest valence-band state as functions of the superlattice period

Following on from this, Fig. 16.10 displays the results of calculations of the energies at the centre of the superlattice Brillouin zone ($k = 0$) for the uppermost valence band and the lowermost conduction band, as a function of the superlattice period (expressed in units of A_0). In both cases, the quantum confinement energy decreases as the superlattice period increases, an effect which was commonly observed in the discussions in Chapters 2 and 3. Note that the energies are given in relation to the top of the valence band of bulk GaAs, which is set to zero, with the bottom of the conduction band coming out as $1.499\,\mathrm{eV}$ in these calculations. One interesting feature is the sudden reduction in the quantum-confinement energy of the conduction-band state when the period decreases to just A_0. In fact, it would be expected *a priori* that the conduction-band minimum would tend towards the band edge of the bulk alloy $Ga_{0.5}Al_{0.5}As$ in the limit of short-period GaAs/AlAs superlattices. By taking a linear combination of the bulk band structures as suggested by the virtual crystal approximation, this would imply an energy of around $2\,\mathrm{eV}$. This behaviour will be discussed in more detail in the next section.

Finally, for this series of calculations, Fig. 16.11 plots the charge density along the centre ($x = y = 0$) of the $n_z = 10$ superlattice spiral, for the uppermost valence-band state. The confinement in the GaAs layers is clear, and, in addition to this, the lowest conduction-band state is also confined in these layers, thus illustrating the Type-I nature of the band alignment. It can be seen from this figure that the wave function does consist of two components, where

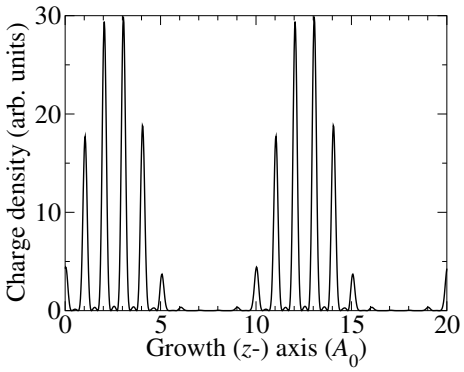


Figure 16.11: The charge density of the uppermost valence-band state along the $x = y = 0$ (z-)axis for two complete superlattice periods

one is rapidly varying and the second is an envelope. This reflects well the idea behind the envelope function approximation, which was discussed (and then used extensively) earlier. The rapidly varying component has a period that is the same as the atomic spacing, and indeed along the axis used in the plot, i.e. the $x = y = 0$ axis, the peaks in the charge density correspond to bond centres. The envelope function is obtained by joining together the peaks in the charge density plot by using the eye, with its form being familiar from Chapter 2.

16.3 Comparison with envelope function approximation

Figure 16.12(a) compares the results of a Kronig–Penney superlattice calculation under the envelope function/effective mass approximations (see Section 2.11), with the empirical pseudo-potential calculations of the lowest-energy conduction-band state as a function of the superlattice period, as previously displayed in Fig. 16.10. It is this comparison that draws attention to the complexity of the data. For guidance, it is necessary to refer to the specialist treatise on empirical pseudo-potential calculations of GaAs/AlAs superlattices from which the pseudo-potentials employed in these calculations are taken (see [1], p. 17396).

Mäder and Zunger indicate that these simple bulk-like potentials are not sufficient to describe the microscopic structure of such a short-period superlattice, and hence the result for $n_z = 1$ could be prone to error. In particular, no account has been taken of the change in the coordination of the interface anions (see Section 16.5). For $2 \leq n_z \leq 8$, the conduction-band minimum for the superlattice originates from the X valleys of the bulk and not from the Γ valley—this is possible because of the indirect nature of AlAs. As the X valley in bulk AlAs lies below that of the Γ valley, then narrow quantum wells can produce such high confinement energies that the eigenstate is influenced by these outlying valleys (in essence!). For $n_z > 8$, the simpler Kronig–Penney model describes the behaviour of the energy state with superlattice period quite well *qualitatively*.

Figure 16.12(b) displays the results of an identical series of calculations for the uppermost (heavy-hole) valence-band state. The effective mass in the AlAs barriers was taken (and fixed) as the bulk value of $0.51m_0$ (Adachi [2], p. 254). In this case, however, the hole effective

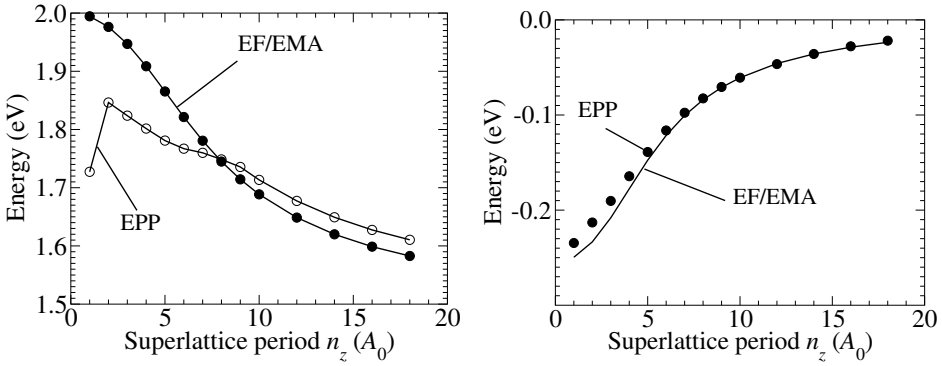


Figure 16.12: Comparison of (a) the lowest-energy conduction-band state and (b) the highest valence-band state given by the empirical pseudo-potential (EPP) calculation in the previous section with that given by the Kronig–Penney model of a superlattice under the envelope function/effective mass approximations (EF/EMA)

mass in the GaAs well regions was used as a parameter and varied in order to produce the best fit to the empirical pseudo-potential results. It can be seen that the Kronig–Penney (envelope function/effective mass approximation) results for an effective mass of $0.45m_0$ fit the pseudo-potential data very well at the larger well and barrier widths, but the match is poorer at narrower widths. This is the result found by Long *et al*. [4], and indeed should be expected, as the envelope function approximation (see Chapter 1) hinges on the point that the wave function can be considered as a product of two components, with one being a rapidly varying Bloch function (which is factorised out) and the second a slowly varying envelope. It stands to reason that, when the period of the envelope approaches that of the Bloch function, as happens here in short-period superlattices, the approximation becomes poorer.

In essence, the result here can be summarised by saying that the effective mass of a particle in a quantum well is a function of the well width. The fact that a constant effective mass predicts too high a confinement energy allows the further deduction that the effective mass must be increased as the well width decreases in order to produce agreement with the empirical pseudo-potential calculations.

In conclusion, the microscopic nature of the pseudo-potential calculation gives more detail and allows for more complexity in the eigenstates of heterostructures than methods based on the envelope function/effective mass approximation. In particular, use of the constant effective mass approximation has been shown to break down for short-period superlattices.

16.4 In-plane dispersion

In addition to computing the dispersion along the line of symmetry, i.e. the growth (z-)axis, the large-basis method, when applied to superlattices as in Section 16.2, can be used to calculate the in-plane (x–y) dispersion. Such knowledge is fundamental for describing the

transport properties of electronic devices, which exploit in-plane transport for their operation, such as high-electron-mobility transistors (HEMTs), as well as optical devices that are influenced by the carriers populating the in-plane momentum states, such as intersubband lasers (see Section 12.2).

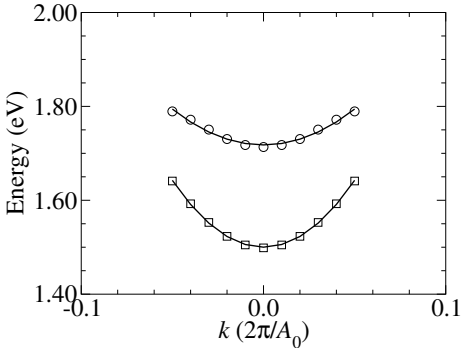


Figure 16.13: The in-plane [100] dispersion curve around the centre of the Brillouin zone for a $(GaAs)_{10}(AlAs)_{10}$ superlattice (top) in comparison with that for bulk (bottom)

Figure 16.13 displays the results of calculations of the in-plane dispersion (top curve), in this case along the [100] direction, for the $(GaAs)_{10}(AlAs)_{10}$ superlattice of the previous section. On the same axes, the bulk dispersion curve (bottom) is also shown for comparison. The subband minima are both measured from the top of the valence band and hence the minimum of the bottom curve represents the band gap of the bulk material, while the difference in the minima represents the quantum confinement energy of this, i.e. the lowest confined state in the superlattice.

Remembering that, for low electron momenta, the E–k dispersion curves are parabolic, then as before the effective mass is given by:

$$m^* = \hbar^2 \left(\frac{\partial^2 E}{\partial k^2} \right)^{-1} \tag{16.15}$$

The fitting of parabolas (the solid curves) to the data points in Fig. 16.13 thus allows the effective masses to be calculated (as before in Chapter 15). Using the potentials of Mäder and Zunger [1], this procedure gives the effective mass along any of the $\langle 100 \rangle$ directions in the bulk crystal as $0.082m_0$. In contrast to this, the in-plane electron effective mass for the superlattice is $0.15m_0$, which is obviously quite different.

This would seem to be a general result, at least for relatively short-period superlattices; note that the period for this example is $10A_0 = 56.5\,\text{Å}$. *A priori*, it might be expected that as there is no confinement in the x–y plane (i.e. parallel to the layers), the dispersion curves would resemble that of the bulk. In fact, as there is confinement along the z-direction, which leads to a shift in the band minimum, the band structure around the minima is clearly different from that of the bulk, the immediate consequence of which is that the effective mass increases.

16.5 Interface coordination

In bulk zinc blende material, each anion, e.g. As in GaAs, is bonded to four Ga cations. It has been shown that the atomic pseudo-potential of this As is different from the As$''$ (say) in AlAs, which is merely a reflection of the different chemical nature of the bonding between Ga atoms and Al atoms arising from their different electronegativities. Although disappointing in that universal atomic potentials cannot be deduced, the calculations thus far have shown how to deal with such a problem.

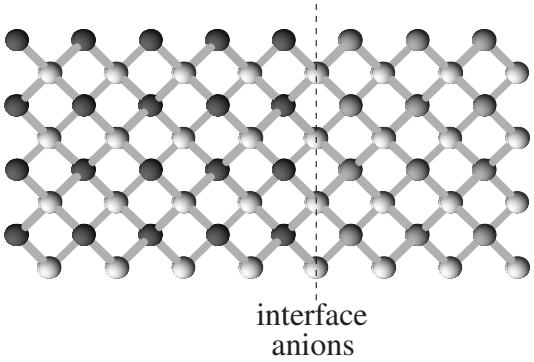


Figure 16.14: The atoms near the heterojunction between two compound semiconductors

Consider now a heterojunction between two compound semiconductors that share a common species, e.g. GaAs/AlAs, as illustrated in Fig. 16.14. The As anion at the interface, as indicated, is bonded to two Ga atoms and two Al atoms, and hence they have neither the character of a fully tetrahedrally coordinated As nor As$''$. In fact, they have a character that is intermediate between the two, which can be described by the mean:

$$\mathrm{As}' = \frac{\mathrm{As} + \mathrm{As}''}{2} \tag{16.16}$$

and hence the interface properties *can* be described better by the mean in the atomic pseudo-potentials, i.e.

$$V_{\mathrm{f}}^{\mathrm{As}'}(q) = \frac{V_{\mathrm{f}}^{\mathrm{As}}(q) + V_{\mathrm{f}}^{\mathrm{As}''}(q)}{2} \tag{16.17}$$

16.6 Strain-layered superlattices

As already mentioned, the more recently deduced pseudo-potentials generally take account of a much greater diversity of experimental information than just the main energy gaps. In particular, one of these additional points is the reproducibility of the deformation potentials from the pseudo-potential calculations (see, for example, [1, 3] for III–V compounds and [5] for SiGe). One consequence of this development is that the pseudo-potentials can be used to accurately describe the microscopic electronic structure in *strain-layered* systems.

Strain arises in heterosystems that are formed from materials with difference lattice constants (see Kelly [6], p. 317, or Adachi [2], p. 271, for an introduction). For example, the lattice constant of GaAs is often quoted as $5.653\,\text{Å}$ and that of AlAs as $5.660\,\text{Å}$, a difference of just 0.1%. Hence, heterosystems formed from these materials (and all of the intermediate $Ga_{1-x}Al_xAs$ alloys) have a quite constant lattice spacing, and for this reason strain in this material system is usually ignored.

The lattice constant of Si is $5.43\,\text{Å}$ and that of Ge $5.66\,\text{Å}$, which represents a lattice mismatch of $(5.66 - 5.43)/5.43 = 4\%$. While this may appears to be a low value, this is not the case for a semiconductor structure (for an introduction to strain in SiGe, see Meyerson [7]). If a layer of Ge is grown on top of a Si substrate, then the first few layers of Ge will assume the in-plane lattice constant of Si, i.e. the Ge atoms will be squeezed beyond their normal equilibrium separation by the crystal potential of Si. To compensate for this, the lattice constant along the growth (z-)axis increases, thus distorting the usual cubic cell into a cuboid. Given the pseudo-potentials now available, it would suffice to calculate the atomic positions, in order to calculate the effect of strain on the electronic structure. One very simplistic model, perhaps suitable for thin strained layers, would be to assume that the material attempts to keep its unit cell at constant volume. Hence, if A_z is the lattice constant along the growth axis then:

$$A_x A_y A_z = A_0^3 \tag{16.18}$$

Therefore, considering Ge on Si, then the in-plane lattice constants of the Ge layer will equate to Si, i.e. $5.43\,\text{Å}$, and hence:

$$A_z = \frac{5.66^3}{5.43^2}\,\text{Å} = 6.15\,\text{Å} \tag{16.19}$$

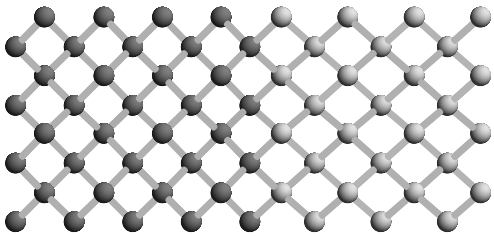


Figure 16.15: The atomic positions at a Si–Ge heterojunction

The application of this model is illustrated for just such a Si–Ge heterojunction in Fig. 16.15—there is an increase in the atomic spacing along the growth (z-)axis as the heterojunction is crossed from the left-hand region of Si atoms to the Ge atoms on the right-hand side. More sophisticated models for the atomic positions in strain-layered semiconductors are available. In particular, Morrison *et al.* [8] compared the results of pseudo-potential calculations of SiGe superlattices, keeping the lattice constant equal to that of a SiGe alloy buffer layer, with those produced with the atomic spacing calculated by minimising the elastic energy [9].

If the strained Ge layer is allowed to grow thicker than the *critical thickness*, then dislocations form that relieve the strain, thus leading to a reduction in the total structural energy of the crystal. Empirical pseudo-potential calculations of the electronic structure around dislocations, and indeed any in-depth study of the particulars of strain-layered superlattices, are beyond the scope of this book. The interested reader is therefore referred to the literature (see, for example, [10, 11] for SiGe, and [12] for $In_xGa_{1-x}As/GaAs$).

The latter of these two material systems is exploited in a relatively common electronic device, namely the pseudomorphic high-electron-mobility transistor, or p-HEMT [13].

16.7 The superlattice as a perturbation

The direct approach to the solution of superlattices by the large-basis method works well, as shown so far; however, this is not always the case. The large number of plane waves required in the expansion generate a very large Hamiltonian matrix, $\mathsf{H}_{\mathbf{G}',\mathbf{G}}$, which in turn requires a computer with a large amount of memory for solution. Such machines are not always available, and it may be necessary to resort to alternative methods. In the case of simple quantum wells and superlattices, one such method is to treat the superlattice as a perturbation on a bulk semiconductor. For the case of quantum dots and wires, an even more powerful computational technique may be required, which will be discussed later in this chapter.

Concentrating on one-dimensional periodic potentials as generated by a pseudo-potential description of a quantum well, the superlattice can be thought of as a perturbation on a bulk 'host' semiconductor [14], as already mentioned. For example, for a $GaAs/Ga_{1-x}Al_xAs$ superlattice, the original host semiconductor may be chosen as GaAs, with the superlattice being formed by exchanging some of the Ga cations for Al. As the word *perturbation* implies, the superlattice wave function is then constructed from a linear combination of the orthogonal bulk eigenfunctions.

Recalling the bulk Schrödinger equation:

$$\mathcal{H}\psi_{n,\mathbf{k}} = E_{n,\mathbf{k}}\psi_{n,\mathbf{k}} \tag{16.20}$$

the introduction of a perturbation, V_{sl}, which is acknowledged to represent the additional potential introduced on formation of the superlattice (sl), yields a new Schrödinger equation with a new set of eigenfunctions:

$$(\mathcal{H} + V_{\text{sl}})\Psi_{N,\xi} = E_{N,\xi}\Psi_{N,\xi} \tag{16.21}$$

where the indices N and ξ respectively represent the principal quantum number and the wave vector of the superlattice eigenfunctions $\Psi_{N,\xi}$. As this is a perturbative approach, it is expected that these eigenfunctions will be expanded in terms of the bulk eigenfunctions, $\psi_{n,\mathbf{k}_\xi}$, i.e.

$$\Psi_{N,\xi} = \sum_{n,\mathbf{k}_\xi} A^N_{n,\mathbf{k}_\xi}\psi_{n,\mathbf{k}_\xi} \tag{16.22}$$

The additional qualification ξ on the wave vectors of the bulk eigenstates indicates that the superlattice eigenstate $\Psi_{N,\xi}$ is summed not just over a particular set of eigenfunctions at a fixed $\mathbf{k}$, but also over the set of bulk wave vectors, which are *zone-folded* into the superlattice Brillouin zone ([15], p. 95). This is illustrated in Fig. 16.16.

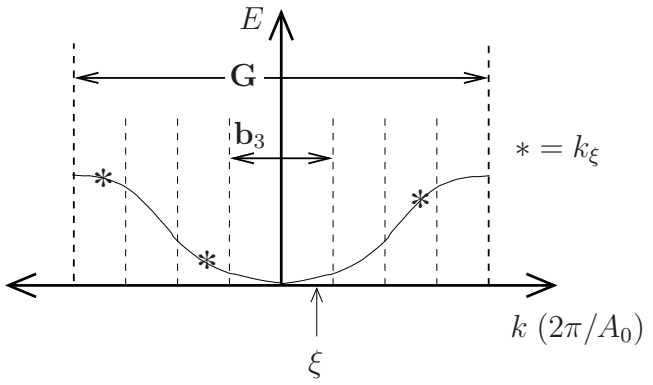


Figure 16.16: Schematic representation of the effect of zone-folding of the bulk Brillouin zone to produce the reduced superlattice mini-zone

The smallest reciprocal lattice vector along one of the $\langle 100 \rangle$ directions, as highlighted by Table 15.1, is the [200] vector. Hence, the von Laue condition, i.e.

$$\mathbf{k}_\bullet \hat{\mathbf{G}} = \frac{1}{2}|\mathbf{G}| \tag{16.23}$$

(see [16], p. 98) is fulfilled for an electron wave vector $\mathbf{k}$ equal to [100], and therefore the edge of the Brillouin zone along the z-axis is at [001], as indicated in the figure. A superlattice, of period $A_{\rm sl} = n_z A_0$ along the z-axis, has its symmetry described by a new set of reciprocal lattice vectors, and indeed the principal primitive is given by:

$$\mathbf{b}_3 = \frac{2\pi}{A_{\rm sl}}\hat{\mathbf{k}} = \frac{1}{n_z}\frac{2\pi}{A_0}\hat{\mathbf{k}} \tag{16.24}$$

which is the same result as that obtained for the simple cubic lattice (see equation (15.112)). Hence, the edge of the superlattice Brillouin zone, which is often referred to as the *mini-zone*, is at:

$$\frac{1}{2}|\mathbf{b}_3| = \frac{2\pi}{2A_{\rm sl}} = \frac{1}{2n_z}\frac{2\pi}{A_0} \tag{16.25}$$

Fig. 16.16 illustrates the mini-zone for a superlattice with period $A_{\rm sl} = 2A_0$, i.e. $n_z = 2$. Thus, the edge of the mini-zone is a quarter of the wave vector that it is for bulk, along this direction. The set of wave vectors $\mathbf{k}_\xi$ are those points within the bulk Brillouin zone that are separated by a superlattice reciprocal lattice vector $\mathbf{b}_3$ from the electron wave vector within the mini-zone. The former are illustrated by the asterisks (*), while the latter is indicated by ξ in the figure.

Therefore, a pseudo-potential calculation of a superlattice of period $A_{\rm sl} = n_z A_0$ requires $N_{\mathbf{k}_\xi} = 2n_z$ wave-vector points from the bulk Brillouin zone. The points themselves are given by:

$$\mathbf{k}_\xi = \xi + (-n_z + i_{\mathbf{k}_\xi})\mathbf{b}_3, \quad i_{\mathbf{k}_\xi} = 0, 1, 2, \ldots, 2n_z - 1 \tag{16.26}$$

and $\mathbf{b}_3$ is given by equation (16.24).

Following standard perturbative theory expansion procedures, substitute $\Psi_{N,\mathbf{k}_\xi}$ from equation (16.22) into the perturbed Schrödinger equation (equation (16.21)), to give:

$$\sum_{n,\mathbf{k}_\xi} A^N_{n,\mathbf{k}_\xi}(\mathcal{H} + V_{\text{sl}})\psi_{n,\mathbf{k}_\xi} = E_{N,\xi}\sum_{n,\mathbf{k}_\xi} A^N_{n,\mathbf{k}_\xi}\psi_{n,\mathbf{k}_\xi} \tag{16.27}$$

Then, using equation (16.20):

$$\sum_{n,\mathbf{k}_\xi} A^N_{n,\mathbf{k}_\xi}(E_{n,\mathbf{k}_\xi} + V_{\text{sl}})\psi_{n,\mathbf{k}_\xi} = E_{N,\xi}\sum_{n,\mathbf{k}_\xi} A^N_{n,\mathbf{k}_\xi}\psi_{n,\mathbf{k}_\xi} \tag{16.28}$$

Multiplying by $\psi^*_{n',\mathbf{k}'_\xi}$ and integrating over all space:

$$\sum_{n,\mathbf{k}_\xi} A^N_{n,\mathbf{k}_\xi}\int \psi^*_{n',\mathbf{k}'_\xi}(E_{n,\mathbf{k}_\xi} + V_{\text{sl}})\psi_{n,\mathbf{k}_\xi}\,\mathrm{d}\tau = E_{N,\xi}\sum_{n,\mathbf{k}_\xi} A^N_{n,\mathbf{k}_\xi}\int \psi^*_{n',\mathbf{k}'_\xi}\psi_{n,\mathbf{k}_\xi}\,\mathrm{d}\tau \tag{16.29}$$

Using the orthonormality of the bulk eigenvectors, i.e. the integral over all space of the product $\psi^*_{n',\mathbf{k}'_\xi}\psi_{n,\mathbf{k}_\xi}$ is non-zero only when $n = n'$ and $\mathbf{k}_\xi = \mathbf{k}'_\xi$, which in mathematical language would be written as:

$$\int \psi^*_{n',\mathbf{k}'_\xi}\psi_{n,\mathbf{k}_\xi}\,\mathrm{d}\tau = \delta_{n'n}\delta_{\mathbf{k}'_\xi,\mathbf{k}_\xi} \tag{16.30}$$

then:

$$\sum_{n,\mathbf{k}_\xi} A^N_{n,\mathbf{k}_\xi}\left(E_{n,\mathbf{k}_\xi}\delta_{n'n}\delta_{\mathbf{k}'_\xi,\mathbf{k}_\xi} + \int \psi^*_{n',\mathbf{k}'_\xi}V_{\text{sl}}\psi_{n,\mathbf{k}_\xi}\,\mathrm{d}\tau\right) = E_{N,\xi}\sum_{n,\mathbf{k}_\xi} A^N_{n,\mathbf{k}_\xi}\delta_{n'n}\delta_{\mathbf{k}'_\xi,\mathbf{k}_\xi} \tag{16.31}$$

Write this as:

$$\sum_{n,\mathbf{k}_\xi} A^N_{n,\mathbf{k}_\xi}H'_{(n',\mathbf{k}'_\xi),(n,\mathbf{k}_\xi)} = E_{N,\xi}\sum_{n,\mathbf{k}_\xi} A^N_{n,\mathbf{k}_\xi}\delta_{n'n}\delta_{\mathbf{k}'_\xi,\mathbf{k}_\xi} \tag{16.32}$$

where the perturbed Hamiltonian matrix elements are given by:

$$H'_{(n',\mathbf{k}'_\xi),(n,\mathbf{k}_\xi)} = E_{n,\mathbf{k}_\xi}\delta_{n'n}\delta_{\mathbf{k}'_\xi,\mathbf{k}_\xi} + \int \psi^*_{n',\mathbf{k}'_\xi}V_{\text{sl}}\psi_{n,\mathbf{k}_\xi}\,\mathrm{d}\tau \tag{16.33}$$

or

$$H'_{(n',\mathbf{k}'_\xi),(n,\mathbf{k}_\xi)} = E_{n,\mathbf{k}_\xi}\delta_{n'n}\delta_{\mathbf{k}'_\xi,\mathbf{k}_\xi} + V' \tag{16.34}$$

where the potential term has been relabelled as follows:

$$V' = \int \psi^*_{n',\mathbf{k}'_\xi}V_{\text{sl}}\psi_{n,\mathbf{k}_\xi}\,\mathrm{d}\tau \tag{16.35}$$

The problem is therefore reduced to finding the eigenvalues, $E_{N,\xi}$, and eigenvectors, $A^N_{n,\mathbf{k}_\xi}$, of the square matrix, $\mathsf{H}'_{(n',\mathbf{k}'_\xi),(n,\mathbf{k}_\xi)}$. This is a direct analogy of the original equation (equation (15.20)), for bulk, and can again be solved by direct diagonalisation. Furthermore,

if N_n bulk energy levels at any one of the $N_{\mathbf{k}_\xi}$ bulk wave vectors are included, then $\mathsf{H}'_{(n',\mathbf{k}'_\xi),(n,\mathbf{k}_\xi)}$ is of the order $N_n N_{\mathbf{k}_\xi}$.

It remains, then, to calculate the matrix elements of H' (for short), and given that the bulk has already been solved, this manifests itself merely as deducing the integral constituting the second term of equation (16.34). Consider substituting the bulk eigenvectors for their plane wave summations, as detailed in the previous chapter, i.e.

$$\psi_{n,\mathbf{k}} = \frac{1}{\sqrt{\Omega}} \sum_{\mathbf{G}} a_{n,\mathbf{k}}(\mathbf{G}) \mathrm{e}^{i(\mathbf{G}+\mathbf{k})\bullet\mathbf{r}} \tag{16.36}$$

Then:

$$V' = \frac{1}{\Omega} \sum_{\mathbf{G}',\mathbf{G}} a^*_{n',\mathbf{k}\xi'}(\mathbf{G}') a_{n,\mathbf{k}\xi}(\mathbf{G}) \int \mathrm{e}^{-i(\mathbf{G}'+\mathbf{k}\xi')\bullet\mathbf{r}} V_{\mathrm{sl}} \mathrm{e}^{i(\mathbf{G}+\mathbf{k}\xi)\bullet\mathbf{r}} \, \mathrm{d}\tau \tag{16.37}$$

$$\therefore V' = \frac{1}{\Omega} \sum_{\mathbf{G}',\mathbf{G}} a^*_{n',\mathbf{k}'_\xi}(\mathbf{G}') a_{n,\mathbf{k}_\xi}(\mathbf{G}) \int \mathrm{e}^{-i(\mathbf{G}'-\mathbf{G}+\mathbf{k}'_\xi-\mathbf{k}_\xi)\bullet\mathbf{r}} V_{\mathrm{sl}} \; d\tau \tag{16.38}$$

The perturbation V_{sl} can be written as a sum of the atomic potentials over all of the sites where atoms have been exchanged. If $V_{a'}$ is the atomic potential of the newly introduced species at a position $\mathbf{r}'_a$, and V_a is the atomic potential of the original species, then the potential due to the creation of the superlattice is just the difference between the two summed over all of the sites, i.e.

$$V_{\mathrm{sl}} = \sum_{\mathbf{r}'_a} [V_{a'}(\mathbf{r} - \mathbf{r}'_a) - V_a(\mathbf{r} - \mathbf{r}'_a)] \tag{16.39}$$

Again, specifying *all* of the atomic positions $\mathbf{r}'_a$ as a combination of a Bravais lattice vector $\mathbf{R}_{\mathrm{sl}}$ and a basis $\mathbf{t}_{\mathrm{sl}}$ (note that the set of basis vectors $\mathbf{t}_{\mathrm{sl}}$ are the position vectors of the exchanged atom positions within a superlattice unit cell), i.e.

$$\mathbf{r}'_a = \mathbf{R}_{\mathrm{sl}} + \mathbf{t}_{\mathrm{sl}} \tag{16.40}$$

then:

$$V_{\mathrm{sl}} = \sum_{\mathbf{R}_{\mathrm{sl}}} \sum_{\mathbf{t}_{\mathrm{sl}}} [V_{a'}(\mathbf{r} - \mathbf{R}_{\mathrm{sl}} - \mathbf{t}_{\mathrm{sl}}) - V_a(\mathbf{r} - \mathbf{R}_{\mathrm{sl}} - \mathbf{t}_{\mathrm{sl}})] \tag{16.41}$$

Consider the vector $\mathbf{G}' - \mathbf{G} + \mathbf{k}'_\xi - \mathbf{k}_\xi = \mathbf{g}$ (say), in equation (16.38). From Fig. 16.16, $\mathbf{k}'_\xi - \mathbf{k}_\xi$ is clearly a reciprocal lattice vector of the superlattice, and in addition, as both $\mathbf{G}'$ and $\mathbf{G}$ are bulk reciprocal lattice vectors, they are also reciprocal lattice vectors of the reduced symmetry system of the superlattice, albeit some large multiple of the primitive superlattice reciprocal lattice vectors. Hence, $\mathbf{g}$ is also a reciprocal lattice vector of the superlattice.

Substituting for $\mathbf{g}$ and V_{sl} from equation (16.41) into equation (16.38) then gives:

$$\begin{aligned} V' = {} & \frac{1}{\Omega} \sum_{\mathbf{G}',\mathbf{G}} a^*_{n',\mathbf{k}'_\xi}(\mathbf{G}') a_{n,\mathbf{k}_\xi}(\mathbf{G}) \\ & \times \int \mathrm{e}^{-i\mathbf{g}\bullet\mathbf{r}} \sum_{\mathbf{R}_{\mathrm{sl}}} \sum_{\mathbf{t}_{\mathrm{sl}}} [V_{a'}(\mathbf{r} - \mathbf{R}_{\mathrm{sl}} - \mathbf{t}_{\mathrm{sl}}) - V_a(\mathbf{r} - \mathbf{R}_{\mathrm{sl}} - t_{\mathrm{sl}})] \; \mathrm{d}\tau \end{aligned} \tag{16.42}$$

Making the transformation $\mathbf{r} - \mathbf{R}_{\mathrm{sl}} - t_{\mathrm{sl}} \longrightarrow \mathbf{r}$, then:

$$V' = \sum_{\mathbf{R}_{\mathrm{sl}}} \mathrm{e}^{-i\mathbf{g}_\bullet \mathbf{R}_{\mathrm{sl}}} \sum_{\mathbf{t}_{\mathrm{sl}}} \mathrm{e}^{-i\mathbf{g}_\bullet \mathbf{t}_{\mathrm{sl}}} \times \frac{1}{\Omega} \sum_{\mathbf{G}',\mathbf{G}} a^*_{n',\mathbf{k}'_\xi}(\mathbf{G}') a_{n,\mathbf{k}_\xi}(\mathbf{G}) \int \mathrm{e}^{-i\mathbf{g}_\bullet \mathbf{r}} \left[V_{a'}(\mathbf{r}) - V_a(\mathbf{r})\right] \, \mathrm{d}\tau \tag{16.43}$$

Now the product of any Bravais lattice vector of the superlattice $\mathbf{R}_{\mathrm{sl}}$ and a reciprocal lattice vector $\mathbf{g}$ is an integral number of 2π, and hence, for N_{sl} superlattice unit cells within the total volume Ω of the crystal:

$$\sum_{\mathbf{R}_{\mathrm{sl}}} \mathrm{e}^{-i2\pi n} = N_{\mathrm{sl}}, \quad n \in \mathbb{Z} \tag{16.44}$$

and:

$$\frac{N_{\mathrm{sl}}}{\Omega} = \frac{1}{\Omega_{\mathrm{sl}}} \tag{16.45}$$

and therefore:

$$V' = \frac{1}{\Omega_{\mathrm{sl}}} \sum_{\mathbf{G}',\mathbf{G}} a^*_{n',\mathbf{k}'_\xi}(\mathbf{G}') a_{n,\mathbf{k}\xi}(\mathbf{G}) \sum_{\mathbf{t}_{\mathrm{sl}}} \mathrm{e}^{-i\mathbf{g}_\bullet \mathbf{t}_{\mathrm{sl}}} \int \mathrm{e}^{-i\mathbf{g}_\bullet \mathbf{r}} \left[V_{a'}(\mathbf{r}) - V_a(\mathbf{r})\right] \, \mathrm{d}\tau \tag{16.46}$$

Note that, as in the bulk calculation, the integral over all space of the atomic potential multiplied by the exponential factor is assigned as the empirical pseudo-potential form factor, i.e.

$$\frac{1}{\Omega_{\mathrm{c}}} \int \mathrm{e}^{-i\mathbf{g}_\bullet \mathbf{r}} V_a(\mathbf{r}) \, \mathrm{d}\tau = V_{\mathrm{f}}^{\mathbf{r}'_a}(g) \tag{16.47}$$

and:

$$\frac{1}{\Omega_{\mathrm{c}}} \int \mathrm{e}^{-i\mathbf{g}_\bullet \mathbf{r}} V_{a'}(\mathbf{r}) \, \mathrm{d}\tau = V_{\mathrm{f}'}^{\mathbf{r}'_a}(g) \tag{16.48}$$

where $g = |\mathbf{g}|$.

As already mentioned, $\mathbf{g}$ is a superlattice reciprocal lattice vector, and as the period of the superlattice is much greater than the period of the bulk crystal, $\mathbf{g}$ is much smaller than the bulk analogy $\mathbf{q}$. Thus, the pseudo-potential form factor $V_{\mathrm{f}}(g)$ needs to be known at intermediate points between the discrete bulk reciprocal lattice vectors—just as in the case of the large-basis calculation.

Therefore, the potential term becomes:

$$V' = \frac{\Omega_{\mathrm{c}}}{\Omega_{\mathrm{sl}}} \sum_{\mathbf{G}',\mathbf{G}} a^*_{n',\mathbf{k}'_\xi}(\mathbf{G}') a_{n,\mathbf{k}_\xi}(\mathbf{G}) \sum_{\mathbf{t}_{\mathrm{sl}}} \mathrm{e}^{-i\mathbf{g}_\bullet \mathbf{t}_{\mathrm{sl}}} \left[V_{\mathrm{f}'}^{\mathbf{r}'_a}(g) - V_{\mathrm{f}}^{\mathbf{r}'_a}(g)\right] \tag{16.49}$$

Hence, the final form of the Hamiltonian matrix elements H' is obtained by substituting the potential matrix elements V' given in equation (16.49) into equation (16.34):

$$H'_{(n',\mathbf{k}'_\xi),(n,\mathbf{k}_\xi)} = E_{n,\mathbf{k}_\xi} \delta_{n'n} \delta_{\mathbf{k}'_\xi,\mathbf{k}_\xi} + \frac{\Omega_{\mathrm{c}}}{\Omega_{\mathrm{sl}}} \sum_{\mathbf{G}',\mathbf{G}} a^*_{n',\mathbf{k}'_\xi}(\mathbf{G}') a_{n,\mathbf{k}_\xi}(\mathbf{G}) \sum_{\mathbf{t}_{\mathrm{sl}}} \mathrm{e}^{-i\mathbf{g}_\bullet \mathbf{t}_{\mathrm{sl}}} \left[V_{\mathrm{f}'}^{\mathbf{r}'_a}(g) - V_{\mathrm{f}}^{\mathbf{r}'_a}(g)\right] \tag{16.50}$$

Recalling the discussion in Section 15.8 that there is no single As potential for all semiconductors, it is important to realise that, when creating a GaAs/$Ga_{1-x}Al_xAs$ superlattice for solution by this perturbative approach, e.g. by substituting atoms onto some of the bulk sites, it is necessary to substitute both Al cations and the As anions associated with AlAs, as these are different from those associated with GaAs.

Equation (16.50) represents the elements of the square matrix $\mathsf{H}'_{(n',\mathbf{k}'_\xi),(n,\mathbf{k}_\xi)}$, the eigenvalues and eigenvectors of which represent, respectively, the energy levels and wave functions of the superlattice, where they are found by direct diagonalisation.

16.8 Application to GaAs/AlAs superlattices

In this section the series of calculations given in Section 16.2 will be revisited and the lowest conduction-band energy calculated for the same range of GaAs/AlAs superlattices, again as a function of period, but this time by using the perturbative approach in order to implement an empirical pseudo-potential calculation rather than the more direct large-basis method.

While the superlattice period controls the number of bulk electron wave vectors, given by the set $\mathbf{k}_\xi$, there is freedom as to the number of *bands* included in the perturbative expansion set. Such a choice is usually made by considering the system of interest and its relation to the bulk band structure. In particular, calculations of the lowest conduction-band state in a superlattice *could* be thought of as being adequately represented by considering just the lowest bulk conduction-band state, whereas the triply degenerate (when excluding spin–orbit coupling) valence band suggests that all three may be needed for the calculation of the uppermost superlattice valence states.

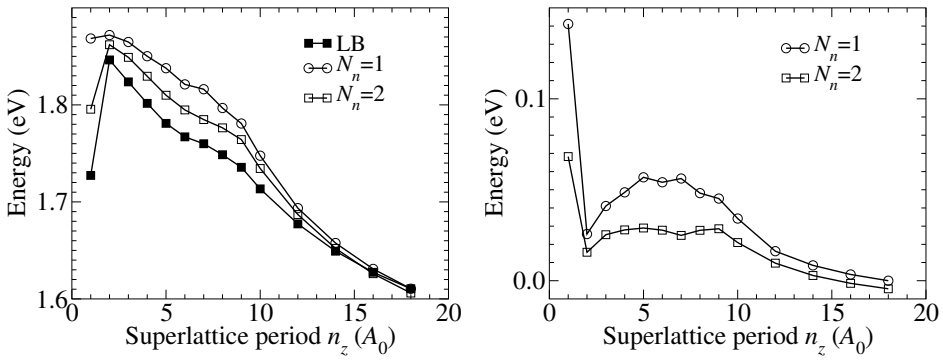


Figure 16.17: Comparison of the perturbative approach to the pseudo-potential calculation of a superlattice, with N_n=1 and 2 conduction band(s), with that of the large-basis (LB) method of earlier. (a) shows the absolute values of the energies, while (b) shows the energy difference between the perturbative and LB models

Figure 16.17a compares the results of calculations obtained by using both the perturbative method and the large-basis method, for the lowest conduction-band state in a GaAs/AlAs superlattice as a function of the period ($n_z(A_0)$). As expected for a perturbative technique, the energy levels obtained are higher than those given by the direct large-basis solution.

It can be seen that some of the details of the data, previously commented upon in Section 16.2, are not reproduced when just one bulk conduction band (N_n=1) is used to provide the expansion set. In particular, the low-energy state of the $(GaAs)_1(AlAs)_1$ short-period superlattice is not reproduced, although there is some evidence of a change in the nature of the states, as evidenced by the functional form of the E–n_z data, at around $n_z = 8$ or 9, as discussed earlier.

When the expansion set is increased to include a second bulk conduction band, the results are much closer to those of the large-basis calculation, as illustrated by the $N_n = 2$ data in Fig. 16.17a. Thus, the obvious conclusion can be drawn that the greater the number of bulk bands included in the expansion set, the better the approximation produced by the perturbative approach. These data are summarised in Fig. 16.17b, which displays the energy difference between the calculations. Analysis of this figure allows a second conclusion to be drawn: the larger the superlattice period, the better the approximation given by the perturbative approach.

Table 16.1 The lowest conduction-band energy level of a $(GaAs)_{10}(AlAs)_{10}$ superlattice as a function of the number of bulk conduction bands N_n included in the basis set

Number of bulk conduction bands, N_n	Lowest energy level, E_1 (eV)
1	1.747636
2	1.734434
3	1.734429
4	1.734421
5	1.734421

Table 16.1 takes this line of inquiry further, displaying the results of calculations of the lowest conduction-band state of the $(GaAs)_{10}(AlAs)_{10}$ superlattice (i.e. $n_z = 10$), as a function of the number of bulk conduction bands (N_n) included in the basis set. The improvement, evidenced by the lowering in the energy level, as implied by the variational principle, is clear when moving from just one bulk conduction band to two. The data show that further improvements are marginal, i.e. less than $1\,\mathrm{meV}$, and are really not worth the additional computation time. The 'best' value offered by this approach, $1.734\,\mathrm{eV}$, compares well with the $1.713\,\mathrm{eV}$ given by the large-basis calculation. The perturbative approach attempts to replace the rather crude set of plane waves with a more sophisticated basis set. However, a discrepancy still exists because, no matter how many bulk bands are included in the superlattice expansion set, each one is still only constructed from the same set of N (in this case, 65) plane waves. This difference between the two approaches diminishes as the superlattice period increases.

The calculations in this section lead to the conclusion that the large-basis approach gives lower, and therefore more accurate (from the variational principle) energy levels. However,

good approximations can be made by using the perturbative approach, particularly for the longer superlattice periods. The issue of computational expense is discussed further in Section 16.11.

16.9 Inclusion of remote bands

It was concluded in the previous section that, for these particular GaAs/AlAs superlattices, if the basis set for the perturbative approach for the superlattice calculation is *limited* to bulk conduction bands, then two bands are sufficient.

However, the question remains: can an extended basis including more remote bands give improved energies? This implies, for the present series of calculations, the necessary inclusion of one or more of the bulk valence bands in the basis set.

Table 16.2 The effect of the inclusion of the uppermost bulk valence band in the basis set of a $(\mathrm{GaAs})_{n_z}(\mathrm{AlAs})_{n_z}$ superlattice

n_z	Energy, E_1^{2CB} (eV)	Energy, $E_1^{\mathrm{2CB+1VB}}$ (eV)
1	1.795507	1.796218
2	1.861993	1.862025
3	1.849019	1.849090
4	1.829507	1.829511
5	1.809960	1.809997
6	1.794834	1.794835
7	1.784877	1.784882
8	1.776331	1.776331
9	1.764226	1.764241
10	1.734433	1.734456
12	1.686980	1.687005
14	1.652077	1.652094

Table 16.2 compares the results of just such calculations, including the uppermost bulk valence-band state, for the series of $(\mathrm{GaAs})_{n_z}(\mathrm{AlAs})_{n_z}$ superlattices considered earlier. It can be seen that the difference between the lowest conduction-band energy level, calculated including the remote valence band (2CB+1VB), and the original (2CB) data is always less than $1\,\mathrm{meV}$, across the entire range of superlattice periods being considered. Therefore, the inclusion of remote bands for these particular superlattices is unnecessary.

16.10 The valence band

As mentioned earlier, the triple degeneracy of the valence band at the centre of the bulk Brillouin zone, produced in this series of calculations without spin–orbit coupling, suggests that all three bands may be needed in any valence-band calculation. Figure 16.18a displays the highest valence-band state of the GaAs/AlAs superlattices, calculated by using the three uppermost bulk valence bands, compared with the results of earlier produced by using

the large-basis method. The same conclusions can be drawn as for the conduction-band

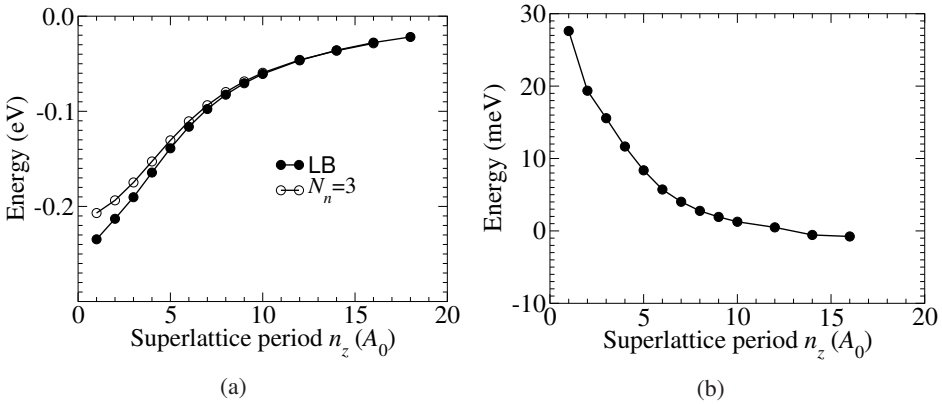


Figure 16.18: Comparison of the perturbative approach to the pseudo-potential calculation of a superlattice, with $N_n = 3$ valence bands, with that of the large-basis (LB) method considered earlier: (a) absolute energy; (b) energy difference between the models

calculations, namely that the perturbative approach does not give energies as low as the large-basis method, while the discrepancy between the two decreases as the superlattice period increases. This is highlighted in Fig. 16.18b.

16.11 Computational effort

The main computational difficulty with the large-basis approach is the demands on computer memory necessary to store the large Hamiltonian matrix. In this respect, the perturbative method can be quite computationally efficient. As mentioned above, the order of the Hamiltonian matrix requiring diagonalisation in the latter method is given by:

$$O(\mathsf{H}'_{(n',\mathbf{k}'_\xi),(n,\mathbf{k}_\xi)}) = N_n N_{\mathbf{k}_\xi} = 2N_n n_z \tag{16.51}$$

where n_z is the period of the superlattice in lattice constants. In comparison, the number of plane waves required in the expansion set for the large-basis calculation was deduced empirically in Section 16.2 as approximately $128n_z$. Hence, for the large-basis approach:

$$O(\mathsf{H}_{\mathbf{G}',\mathbf{G}}) \approx 128 n_z \tag{16.52}$$

Thus, it can be seen that the order of both matrices scales linearly with the superlattice period n_z. However, it is clear that if the number of bands N_n required in the perturbative calculation is less than 64, then the Hamiltonian matrix $\mathsf{H}'_{(n',\mathbf{k}'_\xi),(n,\mathbf{k}_\xi)}$ will be smaller than $\mathsf{H}_{\mathbf{G}',\mathbf{G}}$.

The computational effort associated with the actual process of diagonalisation of these matrices scales as the square of the order, and hence this part of the perturbative approach will require less computer time, as well as less memory, than the large-basis approach.

16.12 Superlattice dispersion and the interminiband laser

As yet, the dispersion curves for electron movement along the growth (z-)axis, i.e. perpendicular transport, have not been calculated by using the perturbative approach to the superlattice.

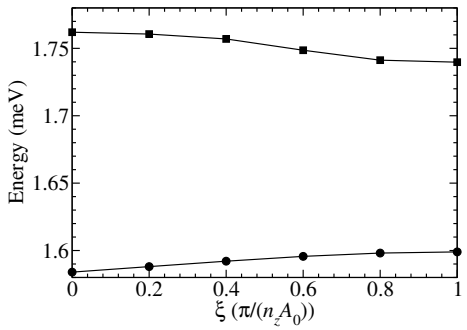


Figure 16.19: The two lowest-energy dispersion curves (minibands) along the growth (z-)axis of a $(GaAs)_{10}(Ga_{0.8}Al_{0.2}As)_{10}$ superlattice

Using the virtual crystal approximation to obtain the form factors for the alloy, Fig. 16.19 plots the two lowest conduction bands of the superlattice, for a range of electron wave vectors ξ across the Brillouin zone of a $(GaAs)_{10}(Ga_{0.8}Al_{0.2}As)_{10}$ superlattice. Just as the superlattice Brillouin zone is often referred to as the *mini-zone*, these superlattice dispersion curves are called *minibands* (see Section 2.11). As discovered in Section 2.11, when using the Kronig–Penney model, the lowest-energy miniband has its minimum at the zone centre, while the second miniband has its minimum at the zone edges.

Thus, if carriers are injected into the upper of these two minibands, then they will rapidly lower their energy, via the fast non-radiative *intraminiband* scattering processes of phonon emission and carrier scattering, as illustrated in Fig. 16.20. These have times similar to those characterising intrasubband processes, perhaps of the order of 100 fs (see Chapter 10). Given the functional form of the E–$\mathbf{k}$ curves within the superlattice Brillouin zone, this energy lowering implies an increasing momentum and the carriers thus move towards the zone edge.

At this point, the carriers have reached the bottom of the miniband and can no longer lower their energy through intraminiband relaxation processes, and they must wait the relatively long time for an *interminiband* scattering event to occur. Following this, any carrier scattering to the states near the zone edge of the lowest miniband will scatter rapidly to lower-momentum states. Such a scenario of a long lifetime in the upper level, compared to a short lifetime in the lower level, for the selection of electron momenta near the superlattice zone edge, is reminiscent of the requirements of the intersubband laser of Chapter 12. Indeed, the combination of these scattering processes within the two lowest minibands of a superlattice forms the basis of a new form of semiconductor laser known as the *interminiband laser* [17, 18]. One exciting feature of this device is that it lases without an overall population

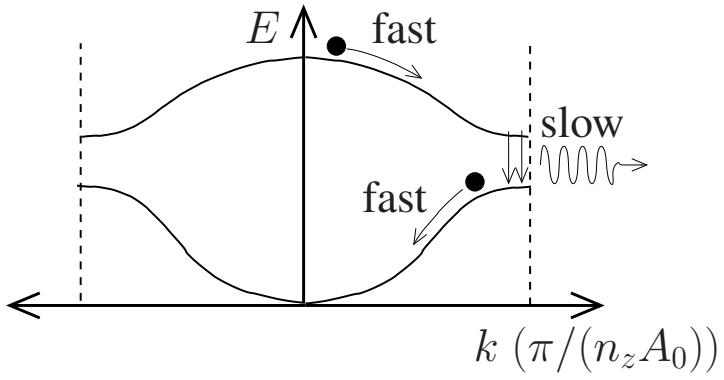


Figure 16.20: The carrier dynamics within an interminiband laser

inversion: it is only necessary to obtain a localised inversion in the states near the zone edge for stimulated emission to occur.

16.13 Addition of electric field

An electric field can be considered merely as a perturbation on the semiconductor system, just as the superlattice itself was. However, given the inherent periodic nature of the pseudo-potential method, the electric field itself has to be periodic and repeated within each unit cell, as shown in Fig. 16.21.

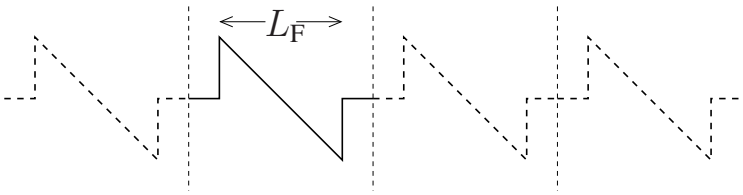


Figure 16.21: The necessary periodic nature of any electric field accounted for with pseudo-potential theory

Taking the first and last atomic positions within the unit cell of interest as the extent of the electric field, and defining the zero-field point as the centre of the cell, i.e. $(z_{N_a} + z_1)/2$, the potential energy due to the electric field would be:

$$V_{\mathrm{F}} = \begin{cases} 0, & z < z_1 \\ -qF\left(z - \frac{z_{N_a}+z_1}{2}\right), & z_1 < z < z_{N_a} \\ 0, & z_{N_a} < z \end{cases} \tag{16.53}$$

where z_1 and z_{N_a} are the z-coordinates of the first (1) and last (N_a) atoms, respectively. With this definition, the field does look like that in Fig. 16.21, with small zero-field regions of width $A_0/2$ in between the regions of linear sloping potential.

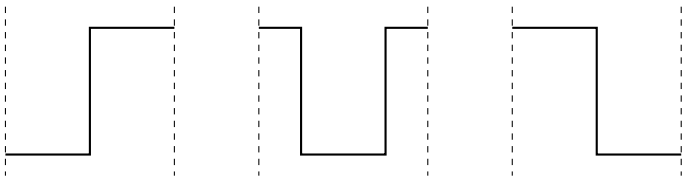


Figure 16.22: The possible unit cells for a pseudo-potential study of a superlattice

The pseudo-potential calculations so far have all centred around true superlattices, i.e. systems of quantum wells with significant overlap between the wave functions of adjacent wells. Single isolated quantum wells can be considered by using pseudo-potential theory, just by making the barriers within each period thick in size, thus producing a large distance between the wells. Hitherto, it was not relevant where the atoms were exchanged within the unit cell to produce a superlattice; for example, as shown in Fig. 16.22, the well could be formed at the beginning of the unit cell, the middle or the end. However, when incorporating an electric field as well, it is important that the field extends either side of the 'region of interest'—this is achieved simply by ensuring the quantum wells are centred in the unit cell.

The effect of an electric field can be calculated for a single quantum well, or for a system of several quantum wells, by the appropriate choice of unit cell (see, for example, Fig. 16.23). Therefore, an electric field can be considered as an additional perturbation with the same

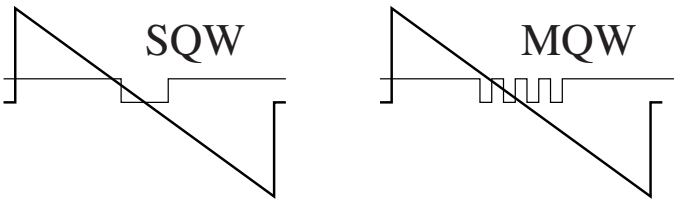


Figure 16.23: The unit cells required to study the effect of an electric field on a single (SQW) or multiple quantum well (MQW)

periodicity as the superlattice unit cell, whether that cell contains one or more quantum wells. Thus, the original Schrödinger equation for the superlattice (equation (16.21)), i.e.

$$(\mathcal{H} + V_{\text{sl}})\Psi_{N,\xi} = E_{N,\xi}\Psi_{N,\xi} \tag{16.54}$$

would have an additional term representing the perturbation due to the electric field, i.e.

$$(\mathcal{H} + V_{\text{sl}} + V_{\text{F}})\Psi_{N,\xi} = E_{N,\xi}\Psi_{N,\xi} \tag{16.55}$$

the consequence of which is that the potential term in equation (16.38) becomes:

$$V'' = \frac{1}{\Omega}\sum_{\mathbf{G}',\mathbf{G}} a^*_{n',\mathbf{k}'_\xi}(\mathbf{G}')a_{n,\mathbf{k}_\xi}(\mathbf{G})\int \mathrm{e}^{-i(\mathbf{G}'-\mathbf{G}+\mathbf{k}'_\xi-\mathbf{k}_\xi)\bullet\mathbf{r}}(V_{\mathrm{sl}}+V_{\mathrm{F}})\ \mathrm{d}\tau \tag{16.56}$$

The manipulation of the first term (V_{sl}) clearly proceeds as before, thus giving the original perturbing potential V' due to the superlattice potential, as defined in equation (16.49), so therefore:

$$V'' = V' + \frac{1}{\Omega}\sum_{\mathbf{G}',\mathbf{G}} a^*_{n',\mathbf{k}'_\xi}(\mathbf{G}')a_{n,\mathbf{k}_\xi}(\mathbf{G})\int \mathrm{e}^{-i(\mathbf{G}'-\mathbf{G}+\mathbf{k}'_\xi-\mathbf{k}_\xi)\bullet\mathbf{r}}V_{\mathrm{F}}\ \mathrm{d}\tau \tag{16.57}$$

Consider just the integral component. Again writing $\mathbf{g} = \mathbf{G}' - \mathbf{G} + \mathbf{k}'_\xi - \mathbf{k}_\xi$:

$$\frac{1}{\Omega}\int \mathrm{e}^{-i(\mathbf{G}'-\mathbf{G}+\mathbf{k}'_\xi-\mathbf{k}_\xi)\bullet\mathbf{r}}V_{\mathrm{F}}\ \mathrm{d}\tau = \frac{1}{\Omega}\int \mathrm{e}^{-i\mathbf{g}\bullet\mathbf{r}}V_{\mathrm{F}}\ \mathrm{d}\tau \tag{16.58}$$

Given the form of V_{F} in equation (16.53) and writing the origin of the electric field potential as $z_0 = (z_{N_a} + z_1)/2$, one obtains:

$$\frac{1}{\Omega}\int \mathrm{e}^{-i\mathbf{g}\bullet\mathbf{r}}V_{\mathrm{F}}\ \mathrm{d}\tau = -\frac{qF}{\Omega}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(z-z_0)\,\mathrm{e}^{-i\mathbf{g}\bullet\mathbf{r}}\ \mathrm{d}z\ \mathrm{d}x\ \mathrm{d}y \tag{16.59}$$

The in-plane (x–y) integrals only have value when the x- and y-components of $\mathbf{g}$ are zero, and are then equal to the length of the crystal in that dimension, i.e.

$$\frac{1}{\Omega}\int \mathrm{e}^{-i\mathbf{g}\bullet\mathbf{r}}V_{\mathrm{F}}\ \mathrm{d}\tau = -\frac{qFL_xL_y}{\Omega}\int_{-\infty}^{+\infty}(z-z_0)\,\mathrm{e}^{-ig_zz}\ \mathrm{d}z\ \delta_{0,g_x}\delta_{0,g_y} \tag{16.60}$$

where the the vector coefficients g_x and g_y are defined by $\mathbf{g} = g_x\hat{\mathbf{i}} + g_y\hat{\mathbf{j}} + g_z\hat{\mathbf{k}}$, and thus the Dirac δ-functions ensure that the integral is non-zero for $\mathbf{g}$ vectors along the axis of the field only. If there are N_{sl_z} unit cells along the z-direction, then:

$$\frac{1}{\Omega}\int \mathrm{e}^{-i\mathbf{g}\bullet\mathbf{r}}V_{\mathrm{F}}\ \mathrm{d}\tau = -\frac{qFL_xL_yN_{\mathrm{sl}_z}}{\Omega}\int_{z_1}^{z_{N_a}}(z-z_0)\,\mathrm{e}^{-ig_zz}\ \mathrm{d}z\ \delta_{0,g_x}\delta_{0,g_y} \tag{16.61}$$

However, the total volume of the crystal $\Omega = L_xL_yN_{\mathrm{sl}_z}n_zA_0$, where, of course, n_zA_0 is the superlattice period. Hence:

$$\frac{1}{\Omega}\int \mathrm{e}^{-i\mathbf{g}\bullet\mathbf{r}}V_{\mathrm{F}}\ \mathrm{d}\tau = -\frac{qF}{n_zA_0}\int_{z_1}^{z_{N_a}}(z-z_0)\,\mathrm{e}^{-ig_zz}\ \mathrm{d}z\ \delta_{0,g_x}\delta_{0,g_y} \tag{16.62}$$

Integrating by parts yields:

$$\begin{aligned}\frac{1}{\Omega}\int \mathrm{e}^{-i\mathbf{g}\bullet\mathbf{r}}V_{\mathrm{F}}\ \mathrm{d}\tau = -\frac{qF}{n_zA_0}\Bigg\{&\left[(z-z_0)\,\frac{\mathrm{e}^{-ig_zz}}{-ig_z}\right]_{z_1}^{z_{N_a}}\\ &-\frac{1}{-ig_z}\int_{z_1}^{z_{N_a}}\mathrm{e}^{-ig_zz}\ \ dz\Bigg\}\,\delta_{0,g_x}\delta_{0,g_y}\end{aligned} \tag{16.63}$$

$$\therefore \frac{1}{\Omega}\int \mathrm{e}^{-i\mathbf{g}_\bullet\mathbf{r}}V_\mathrm{F}\,\mathrm{d}\tau = -\frac{qF}{n_z A_0}\left[(z-z_0)\,\frac{i\mathrm{e}^{-ig_z z}}{g_z}+\frac{\mathrm{e}^{-ig_z z}}{g_z^2}\right]_{z_1}^{z_{N_a}}\delta_{0,g_x}\delta_{0,g_y} \tag{16.64}$$

and:

$$\frac{1}{\Omega}\int \mathrm{e}^{-i\mathbf{g}_\bullet\mathbf{r}}V_\mathrm{F}\,\mathrm{d}\tau = -\frac{qF}{n_z A_0}\left[\left(\frac{i\,(z-z_0)}{g_z}+\frac{1}{g_z^2}\right)\mathrm{e}^{-ig_z z}\right]_{z_1}^{z_{N_a}}\delta_{0,g_x}\delta_{0,g_y} \tag{16.65}$$

which upon evaluation, gives:

$$\frac{1}{\Omega}\int \mathrm{e}^{-i\mathbf{g}_\bullet\mathbf{r}}V_\mathrm{F}\,\mathrm{d}\tau = -\frac{qF}{n_z A_0}\left\{\left(\frac{i\,(z_{N_a}-z_0)}{g_z}+\frac{1}{g_z^2}\right)\mathrm{e}^{-ig_z z_{N_a}} - \left(\frac{i\,(z_1-z_0)}{g_z}+\frac{1}{g_z^2}\right)\mathrm{e}^{-ig_z z_1}\right\}\delta_{0,g_x}\delta_{0,g_y} \tag{16.66}$$

The total potential term for both the superlattice perturbation and the electric field perturbation is obtained by substituting equation (16.66) into equation (16.57), thus giving:

$$V'' = V' + \sum_{\mathbf{G}',\mathbf{G}} a^*_{n',\mathbf{k}'_\xi}(\mathbf{G}')a_{n,\mathbf{k}_\xi}(\mathbf{G})\left(-\frac{qF}{n_z A_0}\right)$$

$$\times\left\{\left(\frac{i\,(z_{N_a}-z_0)}{g_z}+\frac{1}{g_z^2}\right)\mathrm{e}^{-ig_z z_{N_a}} - \left(\frac{i\,(z_1-z_0)}{g_z}+\frac{1}{g_z^2}\right)\mathrm{e}^{-ig_z z_1}\right\}\delta_{0,g_x}\delta_{0,g_y} \tag{16.67}$$

Using the definition for the superlattice perturbing potential V' in equation (16.49), the final form for the Hamiltonian matrix elements *including* an electric field is:

$$H'_{(n',\mathbf{k}'_\xi),(n,\mathbf{k}_\xi)} = E_{n,\mathbf{k}_\xi}\delta_{n'n}\delta_{\mathbf{k}'_\xi,\mathbf{k}_\xi} + \sum_{\mathbf{G}',\mathbf{G}} a^*_{n',\mathbf{k}'_\xi}(\mathbf{G}')a_{n,\mathbf{k}_\xi}(\mathbf{G})$$

$$\times\left\{\frac{\Omega_\mathrm{c}}{\Omega_\mathrm{sl}}\sum_{\mathbf{t}_\mathrm{sl}}\mathrm{e}^{-i\mathbf{g}_\bullet\mathbf{t}_\mathrm{sl}}\left[V_{\mathrm{f}'}^{\mathbf{r}'_a}(g)-V_\mathrm{f}^{\mathbf{r}'_a}(g)\right]-\frac{qF}{n_z A_0}\delta_{0,g_x}\delta_{0,g_y}\right.$$

$$\left.\times\left\{\left(\frac{i\,(z_{N_a}-z_0)}{g_z}+\frac{1}{g_z^2}\right)\mathrm{e}^{-ig_z z_{N_a}} - \left(\frac{i\,(z_1-z_0)}{g_z}+\frac{1}{g_z^2}\right)\mathrm{e}^{-ig_z z_1}\right\}\right\} \tag{16.68}$$

which is an extension of the earlier form in equation (16.50). Inspection of equation (16.68) does raise a small problem, namely divergence of the electric field perturbation when $g_z = 0$. For this particular instance, it is necessary to revisit equation (16.62) and put $g_z = 0$, i.e.

$$\frac{1}{\Omega}\int \mathrm{e}^{-i\mathbf{g}_\bullet\mathbf{r}}V_\mathrm{F}\,\mathrm{d}\tau = -\frac{qF}{n_z A_0}\int_{z_1}^{z_{N_a}}(z-z_0)\;\mathrm{d}z\;\delta_{0,g_x}\delta_{0,g_y} \tag{16.69}$$

and then:

$$\frac{1}{\Omega}\int \mathrm{e}^{-i\mathbf{g}_\bullet\mathbf{r}}V_\mathrm{F}\,\mathrm{d}\tau = -\frac{qF}{n_z A_0}\left[\frac{z^2}{2}-zz_0\right]_{z_1}^{z_{N_a}}\delta_{0,g_x}\delta_{0,g_y} \tag{16.70}$$

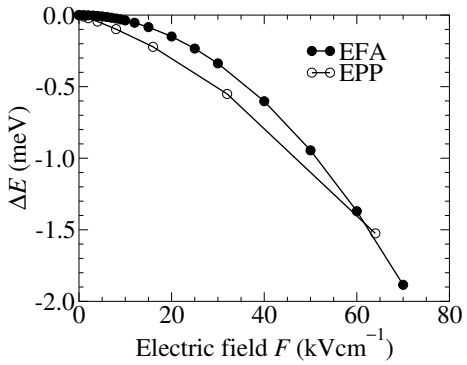


Figure 16.24: The change in the energy of the lowest eigenstate of a 28.25 Å GaAs single quantum well surrounded by $Ga_{0.8}Al_{0.2}As$ barriers, as a function of the applied electric field

Recalling that $z_0 = (z_{N_a} + z_1)/2$, this becomes:

$$\frac{1}{\Omega}\int \mathrm{e}^{-i\mathbf{g}_\bullet\mathbf{r}} V_\mathrm{F}\,\mathrm{d}\tau = 0 \tag{16.71}$$

Therefore, it has been shown that an electric field can be included in the pseudo-potential formalism, with the result being an additional potential term in the Hamiltonian matrix.

Consider the application of an electric field to a single 10 ML (28.25 Å) thick GaAs quantum well surrounded by 20 ML (56.5 Å) thick $Ga_{0.8}Al_{0.2}As$ barriers. Thus, in the periodic formalism characteristic of pseudo-potentials, this would imply a separation of 2×20 ML (113 Å), which should be enough for them to act as independent quantum wells.

Figure 16.24 displays the results of calculations of the change in the energy (ΔE) of the lowest conduction-band energy level as a function of the electric field, as deduced by the empirical pseudo-potential (EPP) method. For comparison, the figure also shows data obtained by the envelope function approximation (EFA) for the same system. The parabolic nature of this, i.e. the quantum-confined Stark effect, is clearly evidenced in the empirical pseudo-potential calculation and, in fact, the energy changes are remarkably similar over the range of electric fields employed. Thus Fig. 16.24 seems to imply, at first sight, that the computationally 'long-winded' approach of the empirical pseudo-potential method is no better than much quicker methods based on the envelope function approximation. Indeed, if Stark effect energy-level changes are all that are of interest, then this would be the case. However, the pseudo-potential method gives much more, e.g. the same calculation can also be used to yield the in-plane dispersion curve, and the pseudo-potential method can account much more fundamentally for valence-band mixing and strain, *at the same time* as modelling the electric field.

16.14 Application of the large-basis method to quantum wires

The perturbative approach to superlattices as described above can be extended to two and three dimensions and hence can be applied to quantum wires and dots. However, the introductory calculations here will be based around the simpler large-basis approach. The unit cell that has to be employed for a quantum wire is just a special case of the extended cube utilised in Section 15.13 for impurities in bulk material. The periodic nature of the pseudo-potential method implies that, in fact, the crystal potential will be an infinite sequence of parallel quantum wires, as shown in Fig. 16.25. However, making use of the symmetry properties of the wire, the latter need only be one lattice constant in length, i.e. assuming the wire to lie along the z-axis, the number of lattice constants in this direction is given by $n_z = 1$ (see Fig. 16.26).

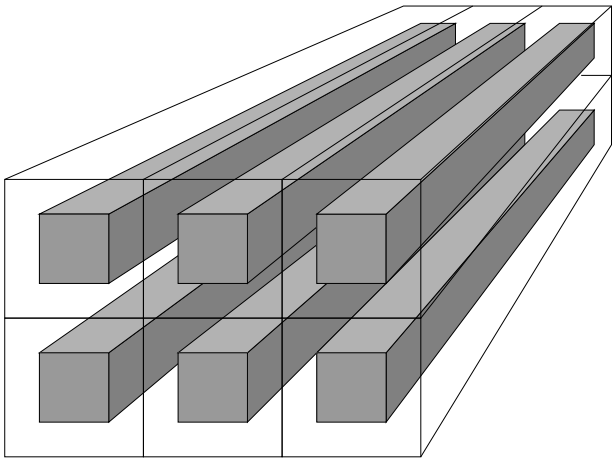


Figure 16.25: The periodic nature of the quantum wire unit cell

A barrier or 'cladding' material surrounds the quantum wire; in this method, it is important that this is of sufficient thickness to isolate the wires from their hypothetical neighbours—just as in the case of modelling single quantum wells using pseudo-potential theory. This thickness depends on the confinement of the carriers within the wire. For example, for wide wires, the confinement is high and hence the overlap of the wave function with that in an adjacent wire is low. Therefore, only a thin barrier layer may be needed.

For the purpose of this illustrative example, consider a square cross-sectional Ge quantum wire surrounded by Si barriers. The atomic positions within the unit cell are illustrated in Fig. 16.27.

The unit cell is just five lattice constants square, i.e. $n_x = 5$, $n_y = 5$, and as mentioned above, just one lattice constant deep, i.e. $n_z = 1$. About the only useful wire that can be accommodated in this box would be three lattice constants square with a single lattice constant barrier, thus giving a total of two lattice constants between the wires; the reason for these limitations on the wire geometry being that, even for this small cross-section wire,

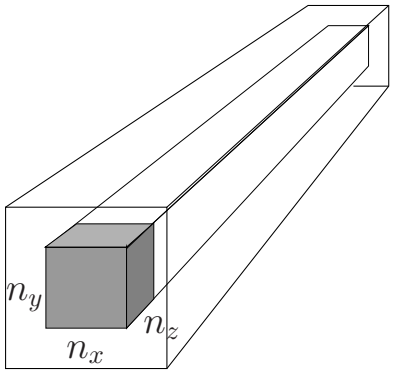


Figure 16.26: The quantum wire unit cell; note the depth of one lattice constant, with sufficient barrier to encompass the wire and, if modelling a single wire, to localise the charge

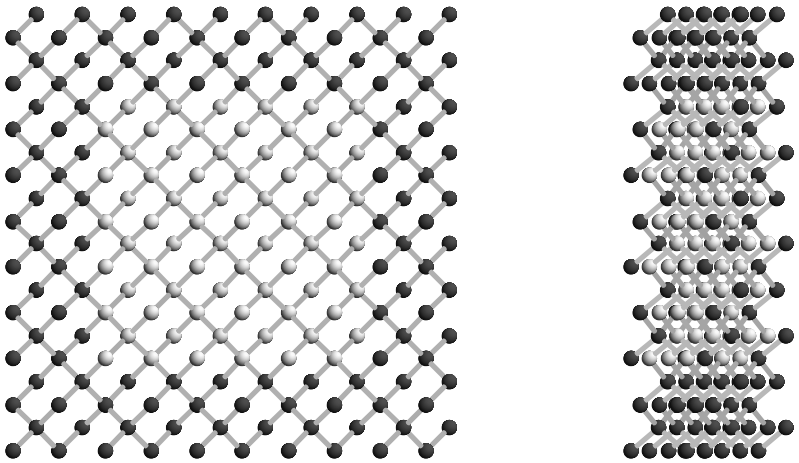


Figure 16.27: The quantum-wire unit cell; note the depth of 1 lattice constant

the number of plane waves required in the expansion would normally be 6625, which would give a Hamiltonian matrix $\mathsf{H}_{\mathbf{G},\mathbf{G}'}$ occupying $6625^2 \times 8 = 351$ MB of computer memory. This can be tackled with a personal computer, but for the purpose of these illustrative calculations, the expansion set will be reduced by truncating at a maximum reciprocal lattice vector of $3 \times 2\pi/A_0$, rather than the usual $4 \times 2\pi/A_0$. This reduces the expansion set to 2751, which still represents 60 MB.

Figure 16.28 illustrates how the untruncated expansion set increases with the number of lattice constants along the side of the wire unit cell. The accuracy of the fit indicates that the expansion set increases as the square of n_x. Given that the number of atoms within the unit cell is also proportional to the area, which in this case of a *square* unit cell is proportional to

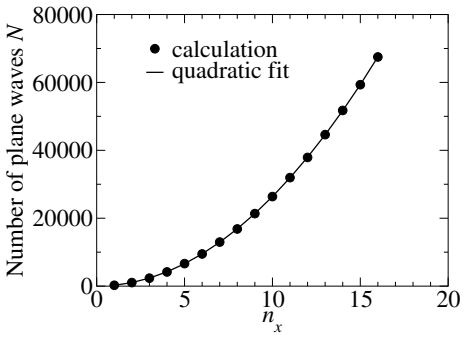


Figure 16.28: The number of plane waves in the expansion set versus the length of the wire side n_x

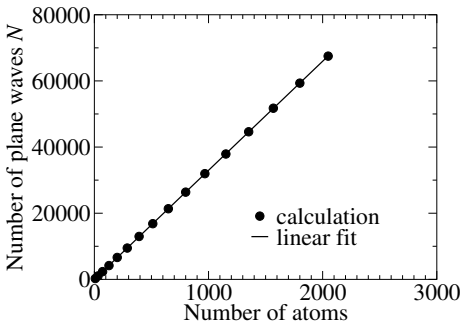


Figure 16.29: The number of plane waves in the expansion set versus the number of atoms in the unit cell

n_x^2, the expansion set may also be expected to be proportional to the number of atoms in the unit cell. This is confirmed by the linear fit shown in Fig. 16.29.

As machine specifications increase, larger wire unit cells will be tractable, thus making this straightforward direct diagonalisation method more useful.

16.15 Confined states

The results of a direct diagonalisation of the Hamiltonian matrix for the quantum wire unit cell in Fig. 16.27 are shown in Fig. 16.30. The latter illustrates the charge density of the lowest conduction-band state for an area the size of the unit cell and across the $z = 0$ plane. The origin of the plot has been shifted slightly in order to centralise the wire within the unit cell.

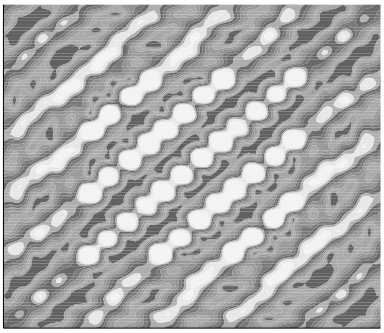
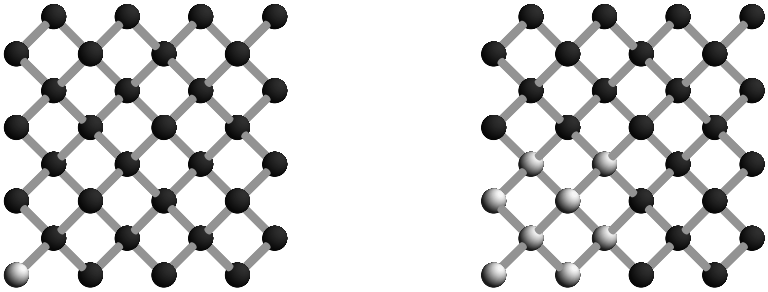


Figure 16.30: The charge density of the lowest conduction-band state over the cross-section of a Ge quantum wire embedded in a Si host



(a) (b)

Figure 16.31: Two of the unit cells for the 'tiny' quantum dot calculations: containing (a) one Ge atom; and (b) eight Ge atoms

16.16 Application of the large-basis method to tiny quantum dots

The problem of the very large basis set is even more acute for a quantum dot, which requires a three-dimensional cubic (or cuboid) unit cell to encase it. Again, with the present computational methods, only small calculations can be attempted, which do, however, illustrate how such calculations are set up.

Consider a unit cell of $2 \times 2 \times 2$ lattice constants; this contains 64 atoms. The (almost ridiculously small) simplest quantum dot that could be placed within this unit cell would be just one atom. Taking silicon as the host material and germanium as the dot, then this particular case is just the isoelectronic impurity dealt with earlier (see Chapter 15). However, in this case, the single Ge atom will be placed at the extreme corner of the unit cell (see Fig. 16.31), which will show the equivalence with the previous example, and will also be the approach used for subsequent calculations. This method of choosing the unit cell is somewhat easier than that used in the quantum wire calculations described earlier in this chapter.

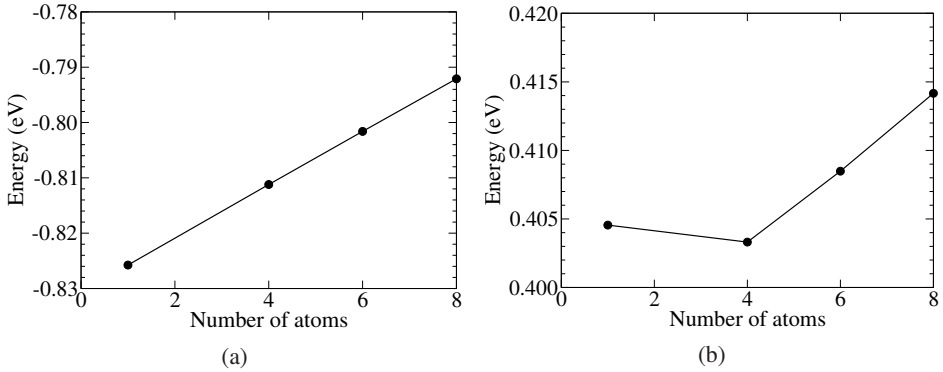


Figure 16.32: The energy of (a) the uppermost valence state and (b) the lowermost conduction state as a function of the number of Ge atoms embedded in a Si host

After choosing the unit cell, which in turn determines the reciprocal lattice vectors in the expansion set, then the empirical pseudo-potential method can yield information such as the energy of the uppermost valence-band state or the lowest conduction-band state. These are illustrated for this simple example in Fig. 16.32.

16.17 Pyramidal quantum dots

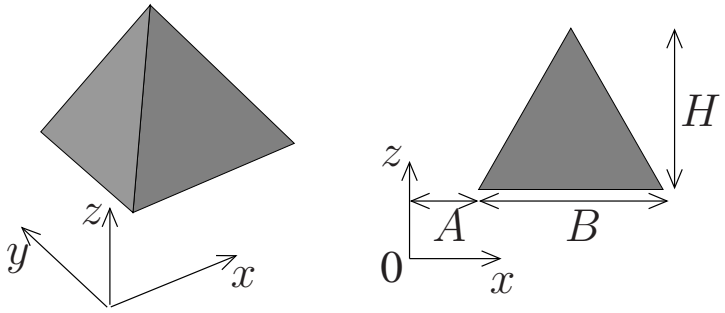


Figure 16.33: Schematic representations of a pyramidal quantum dot

As mentioned in Chapters 8 and 9, the deposition of a thin layer of one material on top of a substrate where there is a large difference in lattice constants can lead to the formation of pyramidal-shaped quantum dots. These quantum dots can be quite large, perhaps a few hundred angstroms across, and are beyond the range of direct diagonalisation of the empirical pseudo-potential Hamiltonian matrix. However, solution has been achieved by using an alternative method (see Section 16.19).

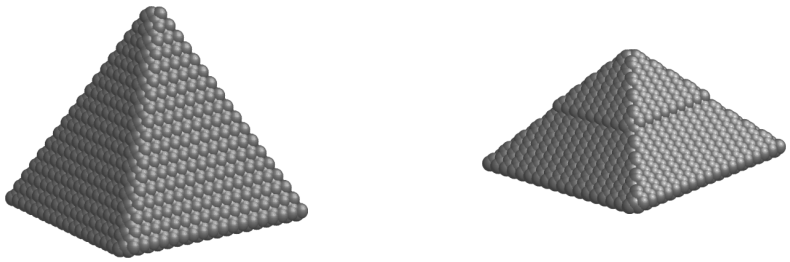


Figure 16.34: The atomic positions for self-assembled pyramidal quantum dots formed from face-centred cubic materials

The setting up of such a calculation requires a knowledge of the atomic positions, so, with this aim, consider the schematic representation of a pyramidal dot shown in Fig. 16.33. Allowing the dot of base length B and height H to be encased within a cubic unit cell of side $A + B + A$, then the equations of the planes represented by the diagonally sloping sides in the right-hand diagram of Fig. 16.33 are as follows:

$$z = \frac{2H}{B}x - \frac{2H}{B} + A \quad \text{and} \quad z = -\frac{2H}{B}x + \frac{2HA}{B} + 2H + A \tag{16.72}$$

Similar equations follow, but this time in terms of y for the remaining two planes. Figure 16.34 illustrates the resulting atomic positions for an elemental quantum dot, such as Ge, as might occur when deposited on a Si substrate (the Si atoms have been removed for clarity). Note that, at certain ratios of the height and base, horizontal ridges form around the circumference of the pyramid.

16.18 Transport through dot arrays

Thus far, only the zone-centre, i.e. $\mathbf{k} = 0$, states have been looked at, but clearly it is possible to calculate transport or other properties that relate to larger momentum states. Transport through the dots is possible, and could be studied by using the empirical pseudo-potential method, although given the necessary periodicity of the technique, solvable problems would be limited to *arrays* of dots.

Figure 16.35 shows the calculated dispersion curves for the uppermost valence-band state and the lowermost conduction-band state for a periodic array of cubic eight-atom Ge quantum dots embedded in a Si host crystal. The dispersion curves along this [001] axis resemble those of a SiGe superlattice, with a zone-folded conduction-band minimum that is nearer the zone centre than in the bulk material; the difference here, though, is that the $[010]$ and $[100]$ directions, i.e. the in-plane directions of the superlattices, also have this same dispersion curve. Thus, the outlying X valleys in the conduction band are brought nearer the zone centre *in all directions*. This contrasts with those superlattices that retain valleys near the zone boundary for electron motion in the conduction band. It may be expected, therefore, that introducing periodic potentials in all directions, as occurs with arrays of quantum dots, might lead to a more direct band gap in SiGe, and hence a better quantum efficiency for light emission.

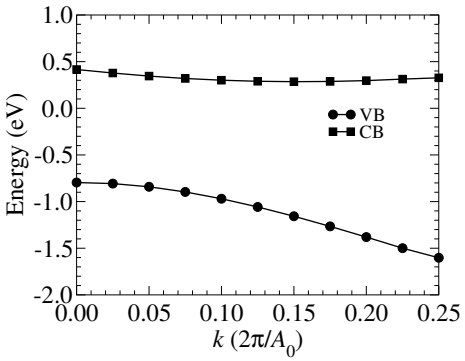


Figure 16.35: Lowest conduction-band (CB) and highest valence-band (VB) dispersion curves through a periodic array of Ge quantum dots embedded in a Si host crystal

16.19 Recent progress

The difficulties associated with applying the techniques outlined in this work to lower-dimensional quantum wires and dots are apparent. To deal with larger quantum dots, remembering that real dots often have a base length in the range 100–400 Å, requires a very large number of plane waves, and indeed the problem becomes too large to handle with present-day computers. The perturbative approach, as derived for the superlatice in Section 16.7, can be extended to deal with quantum dots [19], and this appears to be a promising improvement for modest-sized unit cells.

Zunger and co-workers have demonstrated a new computational technique,which deals with this large-matrix problem. Instead of seeking solutions of the usual Schrödinger equation, i.e.

$$\mathcal{H}\psi_{n,\mathbf{k}} = E_{n,\mathbf{k}}\psi_{n,\mathbf{k}} \tag{16.73}$$

solutions are sought to the alternative expression:

$$\left(\mathcal{H} - E_{\text{ref}}\right)^2 \psi_{n,\mathbf{k}} = \left(E_{n,\mathbf{k}} - E_{\text{ref}}\right)^2 \psi_{n,\mathbf{k}} \tag{16.74}$$

where the reference energy, E_{ref}, can be chosen to lie within the fundamental gap, and hence the valence- and conduction-band-edge states, which are often those of primary interest, are transformed from being arbitrarily high-energy states to being the lowest states. The technique is to minimise the expectation value:

$$\left\langle \psi \left| \left(\mathcal{H} - E_{\text{ref}}\right)^2 \right| \psi \right\rangle \tag{16.75}$$

where the standard empirical pseudo-potential operator of equation (15.5) can be employed. The approach has become known as the *folded spectrum method*; for more details of the theory, see Wang and Zunger [20], while for an example of its application, see Wang *et al*. [21, 22].

Exercises

(1). With computer memory being more readily available, extend the calculations in Section 16.3 to larger superlattice periods. Try several different values for the effective masses of the electron and hole in the envelope function Krönig–Penney (envelope function/effective mass approximation) model and comment on the agreement with the empirical pseudo-potential method.

(2). Implement the extension to include the effect of interface coordination on the anions as described in Section 16.5. Use the short-period GaAs/AlAs superlattice, as described in Section 16.3 (and in exercise (1)) as an example to show the size of the effect on the highest-energy valence- and lowest-energy conduction-band states.

(3). Use the `qwwad_pp_large_basis_so` program to investigate the potential for using a large-basis approach with spin–orbit coupling to perform a short-period superlattice calculation. Choose a superlattice consisting of 10 monolayers of GaAs followed by 10 monolayers of AlAs as the example and calculate the effective mass of the uppermost valence-band state *in the plane* of the quantum wells. Compare with bulk GaAs calculations with and without spin–orbit coupling. Hint: use the GaAs spin–orbit parameters for AlAs.

(4). Use the `qwwad_pp_superlattice` program, which treats a superlattice as a perturbation on the bulk crystal (see Section 16.7), to investigate further the example in Fig. 16.24. In particular, explore the effect on the superlattice energy level as a function of electric field:

   (a) when including a second bulk conduction band in the expansion set for the superlattice;
   (b) for increasing barrier widths;
   (c) for the first excited superlattice state.

(5). The original pseudo-potential calculations of the quantum wire in Section 16.15 used a set of reciprocal lattice vectors that were truncated at $3 \times 2\pi/A_0$. Modern computers have more memory as standard nowadays, so repeat the calculation but with the maximum reciprocal lattice vector at the standard $4 \times 2\pi/A_0$. Compare the energies of the fundamental gap. Comment on the order of the Hamiltonian matrix and the execution time in each case.

(6). Keeping with an eight-atom Ge dot embedded in Si, explore the memory limits of your computer by increasing the size of the Si host crystal from 64 atoms as in Section 16.16. Comment on the memory usage and execution time.

(7). Return to the calculation of the dispersion curve for electron transport through arrays of eight-atom cubic Ge quantum dots embedded in a Si host, as in Section 16.18. Extend/alter the calculations making use of `qwwad_pp_large_basis` to:

   (a) show periodicity in the E–k curve by extending the calculation along the $[001]$ direction to the edge of the second Brillouin zone;
   (b) calculate the dispersion (E–k) along the $[111]$ direction.

References

[1] K. A. Mäder and A. Zunger, 'Empirical atomic pseudopotentials for AlAs/GaAs superlattices, alloys, and nanostructures', *Phs. Rev. B*, **50**:17393, 1994.

[2] S. Adachi, *GaAs and Related Materials: Bulk Semiconducting and Superlattice Properties*, World Scientific, Singapore, 1994.

[3] T. Mattila, L.-W. Wang, and A. Zunger, 'Electronic consequences of lateral composition modulation in semiconductor alloys', *Phys. Rev. B*, **59**(23):15270–15284, 1999.

[4] F. Long, W. E. Hagston, and P. Harrison, 'Breakdown of the envelope function/effective mass approximation in narrow quantum wells', in *The Proceedings of the 23rd International Conference on the Physics of Semiconductors*, Singapore, 1996, pp. 1819–1822, World Scientific.

[5] P. Friedel, M. S. Hybertsen, and M. Schlüter, 'Local empirical pseudopotential approach to the optical properties of Si/Ge superlattices', *Phys. Rev. B*, **39**(11):7974–7977, 1989.

[6] M. J. Kelly, *Low Dimensional Semiconductors: Materials, Physics, Technology, Devices*, Clarendon Press, Oxford, 1995.

[7] B. S. Meyerson, 'High-speed silicon-germanium electronics', *Sci. Am.*, **3**:42, 1994.

[8] I. Morrison, M. Jaros, and K. B. Wong, 'Strain-induced confinement in $Si_{0.75}Ge_{0.25}$ ($Si/Si_{0.5}Ge_{0.5}$) (001) superlattice systems', *Phys. Rev. B*, **35**:9693, 1987.

[9] G. C. Osbourn, 'Strained-layer superlattices from lattice mismatched materials', *J. Appl. Phys.*, **53**:1586, 1982.

[10] J. C. Bean, L. C. Feldman, A. T. Fiory, S. Nakahara, and I. K. Robinson, 'Ge_xSi_{1-x}/Si strained layer superlattices grown by molecular beam epitaxy', *J. Vac. Sci. Technol. A*, **2**:436, 1984.

[11] E. Kasper, H. J. Herzog, and F. Schaffler, 'Si/Ge multilayered structures', in *Physics, Fabrication, and Applications of Multilayer Structures*, P. Dhez and C. Weisbuch, Eds., New York, 1988, vol. 182 of *NATO ASI Series B, Physics*, pp. 229–238, Plenum.

[12] I. J. Fritz, S. T. Picraux, L. R. Dawson, T. J. Drummond, W. D. Laidig, and N. G. Anderson, 'Dependence of critical layer thickness on strain for $In_xGa_{1-x}As$/GaAs strained layer superlattices', *Appl. Phys. Lett.*, **46**:967, 1985.

[13] R. L. Ross, S. P. Svensson, and P. Lugli, Eds., *Pseudomorphic HEMT Technology and Applications*, vol. 309 of *NATO ASI Series E*, Kluwer Academic, Dordrecht, 1996.

[14] M. A. Gell, D. Ninno, M. Jaros, and D. C. Herbert, 'Zone folding, morphogenesis of charge densities, and the role of periodicity in GaAs-$Al_xGa_{1-x}As$ (001) superlattices', *Phys. Rev. B*, **34**:2416, 1986.

[15] M. Jaros, *Physics and Applications of Semiconductor Microstructures*, Clarendon Press, Oxford, 1989.

[16] N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Saunders College Publishing, Philadelphia, 1976.

[17] L. Friedman, R. Soref, and G. Sun, 'Quantum parallel laser: a unipolar superlattice interminiband laser', *IEEE Photonics Technol. Lett.*, **9**(5):593–595, 1997.

[18] G. Scamarcio, F. Capasso, J. Faist, C. Sirtori, D. L. Sivco, A. L. Hutchinson, and A. Y. Cho, 'Tunable interminiband infrared emission in superlattice electron transport', *Appl. Phys. Lett.*, **70**:1796–1798, 1997.

[19] R. J. Turton and M. Jaros, 'Effects of interfacial ordering on the optical properties of Si-Ge superlattices', *Semicond. Sci. Technol.*, **8**:2003, 1993.

[20] L.-W. Wang and A. Zunger, 'Solving Schrödinger's equation around a desired energy: application to silicon quantum dots', *J. Chem. Phys.*, **100**:2394, 1994.

[21] L.-W. Wang, A. Franceschetti, and A. Zunger, 'Million-atom pseudopotential calculation of Γ–X mixing in GaAs/AlAs superlattices and quantum dots', *Phys. Rev. Lett.*, **78**(14):2819–2822, 1997.

[22] L. W. Wang, J. N. Kim, and A. Zunger, 'Electronic structures of [110]-faceted self-assembled pyramidal InAs/GaAs quantum dots', *Phys. Rev. B*, **59**, 1999.

Concluding remarks

A detailed study of this book will have revealed that, although it contains quite a considerable amount of information, there is also a fair bit missing, and indeed there remain one or two question marks over some of the content that has been presented. The reason for this is that the whole area is still a very active field. It is not all ‘done and dusted’, and there still remain some major pieces of theory that need to be developed. It is hoped that this book will provide a stimulus to other workers in the field to fill the gaps—gaps that may well have only become apparent upon reading this summary.

A

Materials parameters

GaAs/$Ga_{1-x}Al_xAs$

- Band gap, $E_g = (1.426 + 1.247x)$ eV
- Band alignment: $\Delta V_{\text{VB}} = 0.33\Delta E_g$; $\Delta V_{\text{CB}} = 0.67\Delta E_g$
- Electron effective mass, $m^* = (0.067 + 0.083x)\, m_0$
- Heavy-hole effective mass, $m^* = (0.62 + 0.14x)\, m_0$
- Lattice constant, $A_0 = 5.65(1 - x) + 5.66x$ Å
- Elastic constants:
  - $C_{11} = 118.8(1 - x) + 120.2x$ GPa
  - $C_{12} = 53.8(1 - x) + 57.0x$ GPa
- Low-frequency (static) dielectric constant $\epsilon_s = 13.18\epsilon_0$
- High-frequency dielectric constant $\epsilon_\infty = 10.89\epsilon_0$
- Material density $\rho = 5317.5\ \text{kg}\,\text{m}^{-3}$
- Longitudinal Optical (LO) phonon energy $E_{\text{LO}} = 36$ meV
- Deformation potential $D_A = 7.0$ eV
- Velocity of sound $v_s = 5117.0\ \text{m}^{-1}$

CdTe/$Cd_{1-x}Mn_xTe$

- Band gap $E_g = (1.606 + 1.587x)$ eV
- Band alignment: $\Delta V_{\text{VB}} = 0.30\Delta E_g$; $\Delta V_{\text{CB}} = 0.70\Delta E_g$
- Electron effective mass, $m^* = (0.11 + 0.067x)\, m_0$
- Heavy-hole effective mass, $m^* = (0.60 + 0.21x + 0.15x^2)\, m_0$

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

$In_{1-x-y}Al_xGa_yAs/AlAs$

- Total band discontinuity, $\Delta V = 2.093x + 0.629y + 0.577x^2 + 0.436y^2 + 1.013xy - 2.0x^2(1 - x - y)$ eV
- Band alignment: $\Delta V_{\text{VB}} = 0.47\Delta E_g$; $\Delta V_{\text{CB}} = 0.53\Delta E_g$
- Electron effective mass, $m^* = (0.0427 + 0.0685x)\, m_0$

$Si_{1-x}Ge_x/Si$

- Lattice constant, $A_0 = 5.431(1 - x) + 5.633x$ Å
- Elastic constants:
  - $C_{11} = 165.8(1 - x) + 128.5x$ GPa
  - $C_{12} = 63.9(1 - x) + 48.3x$ GPa

$GaAs/Ga_{1-x}In_xAs$

- Lattice constant, $A_0 = 5.653(1 - x) + 6.058x$ Å
- Elastic constants:
  - $C_{11} = 118.8(1 - x) + 83.4x$ GPa
  - $C_{12} = 53.8(1 - x) + 45.4x$ GPa

$GaN/Al_xGa_{1-x}N$

- Band gap, $E_g = 3.20(1 - x) + 6.026x$ eV (for $x < 0.45$)
- Conduction-band discontinuity, $\Delta V_{\text{CB}} = 2x$ eV
- Electron effective mass, $m^* = (0.2 + 0.1x)m_0$
- Elastic constants:
  - $C_{11} = 390.0(1 - x) + 396.0x$ GPa
  - $C_{12} = 145.0(1 - x) + 137.0x$ GPa
  - $C_{13} = 106.0(1 - x) + 108.0x$ GPa
  - $C_{33} = 398.0(1 - x) + 373.0x$ GPa
- Piezoelectric constants:
  - $\varepsilon_{13} = -0.49(1 - x) - 0.60x$
  - $\varepsilon_{33} = 0.73(1 - x) + 1.46x$

InAs

- Electron effective mass, $m^* = 0.023m_0$

InP

- Electron effective mass, $m^* = 0.080 m_0$

B

Introduction to the simulation tools

A large part of this book has been concerned with strengthening the connection between the theory of semiconductor heterostructures and their solution by computational methods. To reinforce this latter aspect, the first edition of this book provided a CD-ROM containing simulation software, which allowed all of the worked examples in the book to be reproduced by the reader. The software was subsequently made freely available as an online resource along with the source codes, thus allowing interested readers, should they wish, to repeat the calculations in the book as a learning aid for their own work.

To coincide with this edition of the book, the software has been converted into a free and open-source package, which is hosted online at:[1]

`https://sourceforge.net/projects/qwwad/`

and

`https://launchpad.net/qwwad/`

where the '`qwwad`' is derived from 'quantum wells, wires and dots'. There are a number of important motivations for this change, which benefit both the users and the developers of the code. Some of the most significant benefits are:

- The project website provides a purpose-built location for providing updates, user support and coordinating development of the software.
- Better packaging tools are provided, making it much simpler to install and run the software on selected operating systems.
- The user license has been formalised (under the GNU General Public License v3.0), so that the code can be freely used, distributed and modified, but the developers must be attributed by citing this book, the project website and any relevant papers.
- It is much easier for multiple developers to work on the code. As a consequence, it is much easier for users to contribute their own enhancements and bug fixes to the project!
- A bug tracker is now available, allowing problems to be reported easily by users.

[1]The project may, in future, be relocated to a different host, which can be located through a search for the 'QWWAD' simulation software.

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

The computer codes themselves have been developed over a number of years; note, however, that these are working research codes, which means that they require some degree of expertise in order to run them and to get the most out of them. The codes were developed on UNIX-like systems and the odd one does contain a system call, but as they conform to the ANSI standard, they should execute under other operating systems, with a little effort.

B.1 Documentation and support

This appendix is intended only to give a brief overview of the software, and the main location for documentation is the wiki on the QWWAD project website. There, the user will find tutorials, worked examples, installation instructions and other documentation. It is also important to note that since the software is under active development, this appendix may become outdated and the reader is strongly encouraged to refer to the project website as their principal resource.

It is difficult to guarantee that such an extensive range of source codes will be totally 'bug-free', and the authors therefore welcome any 'bug reports' and 'fixes'. A bug tracker, software updates and further information are available on the project web page. Contributions of new features from interested readers are, of course, very welcome.

B.2 Installation and dependencies

Full installation instructions are given in the wiki on the project website, and are summarised as follows. The simulation programs make extensive use of free numerical libraries (principally Armadillo, Boost, LAPACK and the GNU Scientific Library). These are highly optimised for speed and robustness, and help to improve the quality and readability of the simulation code. These libraries must be present on the user's system in order to compile and run the programs.

An automated installer is provided for Ubuntu Linux, which is a popular, user-friendly and free operating system. Users are strongly recommended to use this approach, since it automatically installs all the necessary libraries, removes the need for the user to compile the code themselves, and benefits from more thorough testing than on other systems.

Alternatively, it is of course possible to download the code from the project website and compile and install it manually. After unpacking the code into a new folder, it can be installed using the standard UNIX-like approach of typing:

```
./configure
make
sudo make install
```

The first command inspects the user's system for all required tools and software libraries and produces an error message if anything is missing. The second command compiles the code and links it to the appropriate libraries. The final command installs the code into the user's system (note that this step requires administrator privileges). The project wiki provides more information about installation to alternative (non-admin) locations, and other options.[2]

[2] At the time of writing, an easier and more flexible installation approach based on the CMake build system is being developed. Full documentation of this feature will be placed on the project website when available.

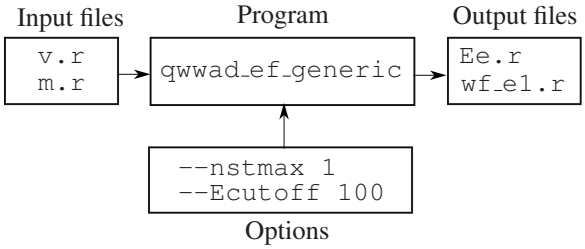


Figure B.1: A typical program call diagram, in which a program (`qwwad_ef_generic`) is run using the `--nstmax 1` and `--Ecutoff 100` options. The program reads the input files `v.r` and `m.r` from the current folder, and writes the output files `Ee.r` and `wf_e1.r`

B.3 Simulation programs

All simulation programs are written in the C or C++ programming languages, although it should be noted that the user is not expected to have any knowledge of programming. The computer codes can be treated as 'black boxes' that need to be fed input data (most of which is generated automatically) and will then give output data. However, it is hoped that by studying the source code in conjunction with this book, the user will gain a sound understanding of the computation methods involved.

There is a philosophy behind the structure of the computer codes. The calculations are, wherever possible, broken down into the smallest possible units, which allows for a great deal of flexibility. For example, the exciton-binding-energy calculation is written for any pair of electron and hole wave functions. Therefore, after these are generated by some solution of a single quantum well, the binding energy in a single quantum well can be calculated. If the binding energy of an exciton in a double quantum well is required, then the user need only calculate the one-particle solutions of the double quantum well, and supply these as input to the exciton-binding-energy calculation. When additional calculations for diffusion and electric field are added *before* the exciton calculation, it becomes apparent that the possible permutations are almost endless.

All of the simulation programs are described in the wiki on the project website, but they are all run using standard commands in a UNIX-like terminal as follows. Each program typically accepts a set of input files and generates a set of output files. For example, the Schrödinger equation can be solved numerically using the `qwwad_ef_generic` program, as illustrated in Fig. B.1. As inputs, this program requires the potential and effective mass profiles at each spatial location in the system. These are stored in plain text files in the current directory called `v.r` and `m.r`, respectively. Each file contains two columns, the first containing a list of spatial locations (in metres) and the second containing the potential (in joules) or the effective mass (in kilograms), respectively.

The program accepts a number of options to control its behaviour. For example, the number of quantum states found by the solver can be limited to one using the `--nstmax 1` option, and the energy range to search for states can be set to 100 meV using the `--Ecutoff 100` option. This would be entered in the UNIX terminal as:

```
qwwad_ef_generic --nstmax 1 --Ecutoff 100
```

Note that options really are considered *optional* here, and a 'sensible' default value will be used by the program if they are not specified. A full list of options for a given program can be seen by running:

```
qwwad_ef_generic --help
```

In many cases, a short form of the option can be used as an alternative. For example, the help text for a program can also be obtained using:

```
qwwad_ef_generic -h
```

The program generates a set of plain-text output files in the current folder, which contain the results of the calculation. In this case, the Schrödinger equation solver generates files called `Ee.r` and `wf_e1.r`, which contain the energy and wave function of the ground state, respectively. In addition to the description of each program on the project wiki, online help can be obtained quickly (including examples and lists of input/output files) from manual pages in the UNIX terminal. For example, the manual for the `qwwad_ef_generic` program can be viewed by typing:

```
man qwwad_ef_generic
```

B.4 Introduction to scripting

The motivation behind this approach was to allow for automation of large numbers of simple calculations and to avoid the need to recompile the programs after the change of a system or material parameter. While each calculation on its own can be trivial and of little use, combining many calculations together to explore, say, the effect of the well width on the ground state energy in a finite square well is of more interest. This could be achieved by using a a short shell script to run the square well solver (`qwwad_ef_square_well`) using a range of different well widths, and to tabulate the result, i.e.

```
#! /bin/sh

# Loop over well width
for L in 20 40 60 80 100 120 140 160 180 200; do

 # Solve the Schroedinger equation
 qwwad_ef_square_well --wellwidth $L

 # Read the energy from file (2nd column)
 E=`cut -f2 Ee.r`

 # Tabulate the length and energy to file
 printf "%e %e\n" $L $E >> E-L.r
done
```

The first line just specifies that this script should run under the standard shell. The loop defines the values which the well width `L` should take. Within the loop itself, which is specified between the `do` and `done` keywords, there are three commands.

- The first calculates the ground state energy of a single quantum well of width `L` using the `qwwad_ef_square_well` program.
- The second uses the standard UNIX `cut` utility to extract the energy from the second column of the `Ee.r` output file, and stores it as the variable `E`.
- The final command appends the well width and the energy of the ground state to the file `E-L.r`, which stands for 'Energy versus L results'.

B.5 Example calculations

The simulation software package installs a folder containing scripts that generate (almost) all the figures in the book. The examples for each chapter are installed into a separate folder, and the scripts needed to generate each figure are tabulated in an `index` file in the folder. For example, the scripts for Chapter 5 are located (by default) in:

```
/usr/share/qwwad/examples/impurities/
```

The `index` file in this folder shows, for example, that one must run the `E-binding-3D.sh` script in order to generate the data shown in Fig. 5.11.

The scripts are run by simply typing `sh`, followed by the name of the script. For Fig. 5.11, the command would be:

```
sh /usr/share/qwwad/examples/impurities/E-binding-3D.sh
```

Note that it is also possible to copy the script into the current directory, meaning that the path to the script need not be specified. On completion, the script generates the `E-binding-3D.dat` file, which can be plotted using a graph-plotting package such as `xmgrace` or `Origin`.

Index

acceptors, 149, 151
acoustic deformation potential scattering, 329
acoustic phonons, 329
adiabatic approximation, 506
AlAs, 4
alloy disorder scattering, 363
    quantum wells, 367
alloys, 522
angular momentum
    orbital, 515
    spin, 515
anion, 519
annealing, 138
anti-crossing, 112
artificial atoms, 279
asymmetric quantum well, 9
atomic basis, 3
atomic form factors, 524
Auger scattering, 354, 355
Auger-type processes, 340, 354

band bending, 125
band filling, 125, 126
band gap, 7, 12, 522
band non-parabolicity, 11, 115
band offset, 536, 548
band structure, 513
band warping, 488
barriers
    double, 66
    single, 64
basis, 510
    vector, 510, 525
beer, 6
BenDaniel–Duke boundary conditions, 46
beryllium, 145
bi-intrasubband scattering, 409
biaxial stress, 225
Bloch functions, 484
Bloch states, 58
Bloch wave function, 483
Bohr magneton, 73
Boltzmann transport equation, 409
Born approximation, 339
bosons, 309
boundary conditions, 105
    Dirichlet, 472
    Neumann, 472
bowing parameter, 523
Bravais lattice, 510
    superlattice, 545
    vector, 525
Brillouin zone, 16, 513, 556
    superlattice, 61
buffer, 234

carrier mobility, 454
carrier–carrier scattering, 339, 344
    computational implementation, 351
    form factor, 342
    intersubband, 350
    intrasubband, 353
    screening, 348
    thermal averaging, 350

Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, Fourth Edition.
Paul Harrison and Alex Valavanis.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

carrier–photon scattering, 379
cation, 519
Caughey–Thomas mobility model, 456
cavity, 384
CdTe, 4
charge density, 519–521, 538, 539
chemical beam epitaxy, 143
chemical bonds, 308
chirped superlattice, 9
classical mechanics, 96
closed-form solution, 81
colloidal dots, 280
complementary error functions, 136
complex band structure, 490
compressive strain, 228
computational considerations, 351, 361
    effective infinities, 201
computer, 536
    memory, 545
conduction band, 6
continuum states, 425
Coulomb potential, 185
covalent, 519
critical thickness, 230, 555
cyclotron frequency, 75

Debye model, 469
Debye temperature, 469
decoupled equations, 252
deformation potentials, 234, 298, 522, 536, 553
    acoustic phonon, 329
    optical phonon, 336
    Pikus–Bir, 492
delta-doping, 144
delta-layers, 144
density matrix, 427
density of states, 26, 125
    bulk, 27
    effective, 191
    one-dimensional, 273
    quantum wires, 273
    three-dimensional, 27
    two-dimensional, 29
diamond, 519, 535
dielectric continuum model, 339
differential equations
    linear second-order, 155
diffraction, 15
diffusion, 131
    boundary conditions, 135
    substitutional-interstitial mechanism, 139
diffusion coefficient, 133
    concentration-dependent, 139
    constant, 137
    depth-dependent, 140
    time-dependent, 142
dilation, 224
diluted magnetic semiconductors, 179
dipole matrix element, 244, 283, 384, 424, 433
direct diagonalisation, 536, 557
dislocations, 555
dispersion, 252, 388–390, 395, 397, 502, 551
    anisotropic, 501, 502
    dispersion curves, 2, 5, 15, 108, 275, 483, 501, 549, 552, 564, 575
    in-plane, 22, 24
    non-parabolic, 11, 24, 115, 488, 501
    parabolic, 2, 16, 250, 274
    phonons, 309, 331
    waveguide, 450
donors, 149
dots, *see* quantum dots
double barrier, 66
    quasi-bound states, 68
    resonance energies, 69
    transmission coefficient, 68
double heterostructure (DH) laser, 458
double quantum well, 9, 105
double-crystal X-ray diffraction, 142
double-precision arithmetic, 536
Drude free-electron model, 454

effective density of states, 191
effective mass, 522
    approximation, 5, 21
    density-of-states, 28

mismatch, 46
effective mass approximation
band non-parabolicity, 178
elastic compliance constants, 225
elastic deformation, 223
elastic scattering, 306
elastic stiffness constants, 225
electric dipole approximation, 396
electric fields, 55, 109
electron temperature, 32, 325, 328, 354, 356
electron–electron scattering, *see* carrier–carrier scattering
electron-beam lithography, 250
electronegative, 553
electrostatics, 117
empirical pseudo-potential theory, *see* pseudo-potential theory, 543, 570
envelope function approximation, 10, 21
envelope functions, 484
epitaxial, 143
error functions, 136
exchange, 340
excitons
1s, 204
2D and 3D limit, 208
2s, 219
binding energy, 139
binding energy in bulk, 198
Bohr radius, 198
high-energy excitation, 197
in infinite quantum wells, 208
multiple quantum wells, 214
resonant excitation, 197
self-consistency, 218
single quantum wells, 212
Stark ladders, 216
superlattices, 214
twin excitons, 211
extrinsic carriers, 119

face-centred cubic, 518
far-infrared, 420, 464
Fermi energy, 27, 35, 349
Fermi wave vector, 27, 348
Fermi's golden rule, 304, 306, 330, 339, 340, 409
two-dimensional systems, 357
Fermi–Dirac, 336
final-state blocking, 308, 345, 350
finite differences, 471
first derivative, 85
matrix methods, 93, 109, 121, 473
second derivative, 86
finite well
unbound states, 44
folded spectrum method, 576
football, 64, 214
form factor, 312, 330
carrier–carrier scattering, 342
Fourier's law, 467

GaAs, 4
charge densities, 521
GaN, 4, 518
Ge
charge densities, 538
growth direction, 549
guided modes, 449
gyromagnetic spin splitting, 73

harmonic oscillator, 96
heat equation, 467
finite-difference approximation, 471
steady-state solution, 472
time-resolved solution, 475
heat flux, 467
heavy hole, 488
Heisenberg's uncertainty principle, 50
HEMT, 124, 240, 552
hermiticity, 494
heterojunction, 7, 131
heterostructure, 8
heterostructure field-effect transistor, 124
hexagonal close-packed, 518
HFET, 124
HgTe, 4
high-electron-mobility transistor, 124, 240, 552

hole–hole scattering, *see* carrier–carrier scattering
holes, 5, 149
homojunction, 131
Hooke's law, 225
hydrodynamic model, 339
hydrogen molecule, 106

impurities, 152, 537
  binding energy, 149
  Bohr radius, 150
  central cell correction, 175
  excited states, 178, 183, 187
  isoelectronic, 533
  Schrödinger equation, 152
  screening, 176
  screening length, 176
impurity scattering, 359
  screening, 362
$In_xGa_{1-x}N$, 4
$In_{1-x-y}Al_xGa_yAs$, 116
independent electron approximation, 506
independently thermalised, 354, 356
infinite well, 19, 24
InGaAs, 555
InP, 4
integration
  by parts, 157, 163, 170, 172
  by substitution, 159, 160
interband, 112, 379
interface, 553
interface mixing, 131
interface roughness scattering, 368
  quantum wells, 372
interminiband, 564
interminiband laser, 61, 564
interstitial, 139
intersubband, 313, 323, 335, 380, 564
  absorption, 383, 386
  Auger-type processes, 354
  devices, 354
  emitters, 386
  lasers, 116, 386, 407, 552, 564
  optical nonlinearities, 394
intraminiband, 564
intrasubband, 312, 324, 335, 340, 353, 564
  Auger-type processes, 355
inverse screening length, 328
ion implantation, 140
ionicity, 519
ionisation energy, 117
isoelectronic impurities, 533, 537

Joule heating, 466

k.p theory, 483
kinetic energy operator, 19, 177

Landé factor, 73
laser
  double heterostructure, 458, 460
  mirror loss, 461
  threshold gain, 461
  unipolar, 456
laser diodes, 407, 441
lattice constant, 554
lattice matching, 232
lattice mismatch, 554
lattice vectors, 3
  primitive, 3
Legendre polynomials, 269
light hole, 488
light-emitting diodes, 407
linear momentum operator, 19
Liouville equation, 430
LO phonon, *see* longitudinal optic phonon
longitudinal optic phonon, 310
  cut-off, 323, 326
  scattering, 311
  screening, 328
Lorentzian line shape, 389

magnetic field vector potential, 73
magnetic flux density, 73
Maxwell's equations, 443
microcavity, 384
mid-infrared, 462
mini-zone, 556, 564
miniband, 61, 564

width, 62
mirror loss, 461
misfit dislocations, 232
mobility, 63
MODFET, 124
modulated-doped field-effect transistor, 124
modulation doping, 117, 124
molecular beam epitaxy, 143
Monte Carlo simulation, 133, 141, 426
multiband effective mass method, 484
multiband envelope function, 484
multiple quantum well, 108
superlattices, 62

nanocrystals, 280
nearly free electron model, 16
Newton–Raphson iteration, 53
non-equilibrium distributions, 325, 354
non-parabolicity, 11, 24, 28, 34, 115, 488
nonlinear optics, 394
nonlinear susceptibilities, 395
numerical differentiation, *see* finite differences
numerical simulation, 134
numerical stability, 137, 475

optical deformation potential scattering, 336
optical rectification, 397
optical waveguides, 441
optically pumped intersubband lasers, 340
orbital angular momentum, 515
orthonormality, 95, 112

p type, 151
p-HEMT, 555
parabolic potential well, 96
finite, 98
parity, 95, 112
particle in a box, 19
Pauli exclusion, 340, 350
permittivity, 109
perpendicular transport, 564
perturbation theory
first-order, 55
second-order, 56
phonon, 308
acoustic, 329, 335
confined, 338
interaction term, 310
interface modes, 339
modes, 339
optical branch, 337
photolithography, 250
photoluminescence, 92, 139
excitation, 139
piezoelectric polarisation, 238
piezoelectricity, 237
plasma damping frequency, 454
plasma frequency, 454
point defects, 539
Poisson effect, 228
Poisson's equation, 118, 145
self-consistent solution, 145
Poisson's ratio, 228
polarisation, 396
polarisation factor, 348
population ratio, 409
Pöschl–Teller potential, 100
principle quantum numbers, 255
pseudo-potential
atomic, 553
form factor, *see also* atomic form factors, 559
pseudo-potential theory
electric fields, 565
empirical, 505, 543, 570
large-cell calculations, 527
single/multiple quantum wells, 566
spin–orbit coupling, 514, 528
superlattice as perturbation, 555
superlattices, 546
pyramidal dots, 270

quantum boxes, *see* quantum dots
quantum cascade lasers, 234, 407, 461
terahertz, 420
quantum dot infrared photodetectors, 426
quantum dots, 249, 279

arbitrary shape, 282
cuboid, 263, 281
density of states, 273
empirical pseudo-potential calculations, 573
empirical pseudo-potential theory, 525
finite barriers, 264
pyramidal, 250
self-assembled, 250, 271, 292
spherical, 264
strain, 271
sublevels, 263
vertically aligned, 273
quantum mechanical tunnelling, 64
quantum well infrared photodetectors, 234, 425
quantum wells
empirical pseudo-potential theory, 525
finite, 38, 44, 46
infinite, 19
intersubband absorption, 386
optical properties, 379
quantum wires, 249
circular cross-section, 260
density of states, 273
empirical pseudo-potential calculations, 570
empirical pseudo-potential theory, 525
finite barriers, 255, 260
infinitely deep, 253
V-grooved, 250
quantum-confined Stark effect, 569
quantum-confinement energy, 549
quasi-bound states, 68

rare earth, 140
reciprocal lattice, 13
reciprocal lattice vectors, 13
primitive, 13
refractive index, 449
remote bands, 562
research opportunities
excitons 2s, 219
resonance energies, 69
resonant continuum states, 45
resonant tunnelling, 70
resonant tunnelling diode, 70
Reststrahlen region, 453
root finding
bisection, 81
Brent algorithm, 84
Newton–Raphson, 83
secant method, 84
rotations, 223

scattering, 303, *see* longitudinal optic phonon, *see* carrier–carrier scattering
acoustic deformation potential, 329
alloy disorder, 363
carrier–photon, 379
impurity, 359
interface roughness, 368
optical deformation potential, 336
spontaneous emission, 384
two-dimensional electrons, 311
scattering matrix element, 305
Schrödinger equation
with electric field, 55
screen, 348
second harmonic generation, 395
secondary ion mass spectroscopy, 139
selection rules, 95, 112, 380
self-assembled quantum dots, 271, 273
self-assembly, 270
shear strain, 492
shooting method, 87
Si
charge densities, 520, 538, 539
SiGe, 554
SIMS, 139
simulation
diffusion, 134
single barrier, 64
single quantum well, 8
sound wave, 331
sp^3 hybridisation, 514

space charge, 109
specific heat capacity, 467, 469
spherical harmonics, 269
spin angular momentum, 515
spin–orbit coupling, 513, 514, 560, 562
spontaneous emission, 384
standard boundary conditions, 39
Stark effect, 56, 111
    quantum-confined, 569
Stark ladders, 216
state function, 1
stationary states, 40, 54, 68
stepped quantum well, 9
stimulated emission, 409
stochastic, 426
strain, 223, 250, 270, 492, 554
    balancing, 232
    compressive, 228, 492
    hydrostatic, 492
    self-assembled quantum dots, 271
    shear, 492
    tensile, 228, 492
    uniaxial, 492
strain-balancing condition, 232
strain-layered, 553
Stranski–Krastanov, 270
stress, 223
    biaxial, 225
stretches, 223
subbands, 24, 312
    anti-crossing, 112
    equilibrium populations, 32
    non-equilibrium populations, 32
    populations, 32
    thermalised distributions, 36
sublevels, 263
superlattice, 58, 543
    Bravais lattice vectors, 545
    Brillouin zone, 61
    finite, 58, 108
    infinite, 108
    Kronig–Penney model, 58, 108
    multiple quantum wells, 62
susceptibility, 397
symmetry, 95, 112
    translational, 519

TEGFET, 124
tensile strain, 228
terahertz, 420
ternary, 523
thermal backfilling, 466
thermal conductivity, 467, 468
thermal equilibrium, 36, 354
thermal leakage, 466
thermal modelling, 466
thermal resistance, 466
thermalisation, 409
thermalised
    independently, 354, 356
threshold current, 425
threshold gain, 461
transfer matrix technique, 53, 54, 67, 446, 496
transmission coefficient, 64
transport, *see* band structure, carrier scattering
    density matrix, *see* density matrix
transverse electric (TE) modes, 445
transverse magnetic (TM) modes, 445
triangular well, 125
TRIM codes, 141
tunnelling, 64
two-dimensional
    electron gas, 29, 124
    electron gas field-effect transistor, 124
    states, 24
    systems, 75
Type-I, 10
Type-II, 10

uncertainty principle, *see* Heisenberg's uncertainty principle, 113
unipolar lasers, 456

V-grooved quantum wires, 250
vacancy, 539
    charge density, 539
valence band, 5, 483, 515
    charge density, 520, 521, 538, 539

electrons, 514
variational principle, 153, 160, 162, 167, 169, 561
trial wave function, 184
variational parameters, 153, 161, 169
virtual crystal approximation, 523, 549, 564
Voigt's notation, 225
von Laue condition, 556
Von Neumann equation, 430

wave function, 1
wave–particle duality, 3
waveguides
cavity, 459
effective mode index, 449
far-infrared, 464
guided modes, 449
mid-infrared, 462
optical, 441
planar, 441
quantum cascade, 461
transverse electric modes, 445
transverse magnetic modes, 445
wires, *see* quantum wires
wurtzite, 4, 518

Zeeman effect, 179
Zeeman splittings, 180
zero-dimensional systems, 279
zero-point energy, 98, 308
zero-stress, 232
zinc blende, 4, 518, 519, 535
ZnS, 4
zone-folded, 555